

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

THIS PAGE BLANK (USPTO)

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-051889

(43)Date of publication of application : 20.02.1998

(51)Int.Cl.

H04R 3/00

H03H 17/00

H03H 21/00

(21)Application number : 08-206210

(71)Applicant : TOSHIBA CORP

(22)Date of filing : 05.08.1996

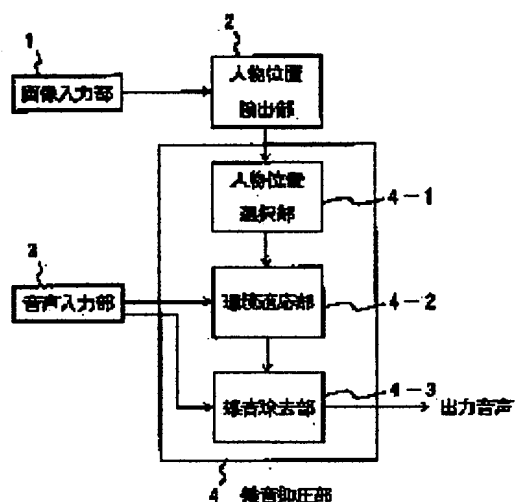
(72)Inventor : NAGATA HITOSHI

(54) DEVICE AND METHOD FOR GATHERING SOUND

(57)Abstract:

PROBLEM TO BE SOLVED: To extract all sounds from person's positions or only the sound from a specific person's position by determining a filter coefficient according to a person's position to be processed which is selected out of detected person's positions or a signal regarding the specific position and a sound source signal.

SOLUTION: A person's position detection part 2 detects the position of a person from an image inputted from an image input part 1. A noise suppression part 4 selects the person's position to be processed out of detected person's positions by a person's position selection part 4-1 and generates the learning signal of an adaptive filter according to the person's position selected by an environment adaptation part 4-2 to determine the coefficient of the adaptive filter. A noise removal part 4-3 filters an input sound from a sound input part 3 by using the determined filter coefficient to suppress its noise. Consequently, when plural persons speak at the same time, all the sounds of the persons can be extracted while having their background noise suppressed or the sound of one person can be extracted while the sounds of other persons are suppressed.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japan Patent Office

UNITED STATES PATENT AND TRADEMARK OFFICE

Application No. 09/111,111
Filed 04/11/2000

Examiner
Patent Attorney

Date

Re: [Illegible]
[Illegible]

[Illegible]
[Illegible]

[Illegible]

THIS PAGE BLANK (USPTO)

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

[Illegible]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-51889

(43) 公開日 平成10年(1998) 2月20日

(51) Int.Cl. ⁹	識別記号	庁内整理番号	F I	技術表示箇所
H 0 4 R 3/00	3 2 0		H 0 4 R 3/00	3 2 0
H 0 3 H 17/00	6 0 1	9274-5 J	H 0 3 H 17/00	6 0 1 G
21/00		9274-5 J	21/00	

審査請求 未請求 請求項の数10 O L (全 26 頁)

(21) 出願番号 特願平8-206210

(22) 出願日 平成8年(1996) 8月5日

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72) 発明者 永田 仁史

大阪府大阪市北区大淀中1丁目1番30号

株式会社東芝関西支社内

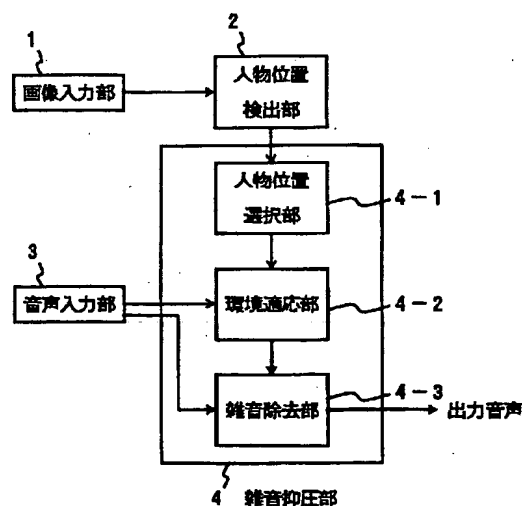
(74) 代理人 弁理士 鈴江 武彦 (外6名)

(54) 【発明の名称】 音声収集装置及び音声収集方法

(57) 【要約】

【課題】 複数の人物位置からの音声に対して、背景雑音を抑えてすべての音声を同時に抽出するがあるいは、特定の人物位置からの音声のみを抽出する。

【解決手段】 入力された画像情報を処理して複数の人物位置を求める人物位置検出部2と、検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択部4-1と、任意に発生された音源信号を、人物位置選択部4-1によって選択された人物位置に配置したものとしたときに観測して得られる入力信号と、選択された人物位置からのすべての音声に対する感度を同時に高くするモードと、選択された人物位置のうち、特定の目的位置からの音声のみを高くするモードのうちいずれかの選択に応じて音源信号から生成される応答信号とに基づいて、フィルタ係数を決定する環境適応部4-2と、決定されたフィルタ係数を用いて、入力された所望の人物位置からの音声のみを抽出する雑音除去部4-3とを具備する。



【特許請求の範囲】

【請求項1】 複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、
複数のチャンネルを介して個々に音声を入力する音声入力手段と、

前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、
この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、

任意に生成した音源信号を、前記人物位置選択手段によって選択された人物位置に配置したものとしたときに観測して得られる第1の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第2の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定手段と、

このフィルタ係数決定手段によって決定されたフィルタ係数を用いて、前記音声入力手段によって入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出手段と、

を具備することを特徴とする音声収集装置。

【請求項2】 前記選択された人物位置のうち、前記特定の目的位置からの音声のみを高くするモードにおいて、複数の目的位置に対応して前記フィルタ係数決定手段及び音声抽出手段を複数個設け、複数の人物位置からの音声を分離して抽出するようにしたことを特徴とする請求項1記載の音声収集装置。

【請求項3】 テスト発声データの入力と前記音声入力手段を介して入力される通常の音声入力の切り替えを指示する入力モード切り替え手段と、入力モードがテスト発声データ入力であるときに、取り込んだテスト発声データのレベルを求めるテスト発声レベル計算手段とをさらに具備することを特徴とする請求項1または2記載の音声収集装置。

【請求項4】 前記画像入力手段によって入力された画像から人物の発声動作に関する情報を位置別に検出する位置別発声動作情報検出手段をさらに具備し、前記フィルタ係数決定手段は、検出した位置別の発声動作に関する情報と、入力された音声から求めた位置別到来パワーの少なくとも一方に基づいて、前記第1の信号である入力信号と前記第2の信号である希望応答信号とを生成することを特徴とする請求項1乃至3のいずれかに記載の音声収集装置。

【請求項5】 複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、
複数のチャンネルを介して個々に音声を入力する音声入力

手段と、

前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、

この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、

この人物位置選択手段によって選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定手段と、

このフィルタ制約設定手段の制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力手段によって入力される音声にフィルタ処理を施して音声を抽出する音声抽出手段と、

を具備することを特徴とする音声収集装置。

【請求項6】 前記フィルタ制約設定手段は、前記選択された人物位置の数が複数の場合に、この複数の人物位置の中の一つの位置を目的位置とし、該目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする第1の制約と、前記目的位置以外の人物位置からの音声に対しては、選択されなかった人物位置と比較して感度を低くする第2の制約をフィルタ処理の制約として設定し、前記音声抽出手段は、前記第1、第2の制約の基にフィルタ出力を最小化してフィルタ係数を決定することを特徴とする請求項5記載の音声収集装置。

【請求項7】 複数の人物を撮影して得られた画像を入力する画像入力手段と、

この画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、

この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、

複数のチャンネルを介して個々に音声を入力する音声入力手段と、

前記人物位置選択手段によって選択された少なくとも一つの人物位置の中の一つの位置を目的位置とし、この目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする制約を設定するフィルタ制約設定手段と、

任意に作成した音源信号を、前記目的位置以外の人物位置に配置したものとしたときに観測される信号を生成する入力信号生成手段と、

前記制約のもとで前記入力信号に基づき目的位置以外の人物位置からの音声に対して感度を低くするようにフィルタを決定するフィルタ決定手段と、

このフィルタ決定手段によって求められたフィルタ係数を用いて、前記音声入力手段によって入力された音声にフィルタ処理を施して音声を抽出する音声抽出手段と、

を具備することを特徴とする音声収集装置。

【請求項8】 前記フィルタ制約設定手段は、前記選択された人物位置の中から複数の目的位置を設定した場合に、この複数の目的位置の一つからの音声に対する感度を、選択されなかった人物位置と比較して高くする制約をフィルタ処理の制約として設定し、前記目的位置以外の人物位置に音源があるものとしたときに観測される入力信号に基づき、前記目的位置以外の人物位置からの音声に対しては感度を、選択されなかった人物位置と比較して低くするようにフィルタを設定するフィルタ設定手段と音声抽出手段とを、前記目的位置の変更に対応して複数の個設け、複数の人物位置からの音声を分離して抽出するようにしたことを特徴とする請求項7記載の音声収集装置。

【請求項9】 複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、複数のチャンネルを介して個々に音声を入力する音声入力工程と、前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、任意に生成した音源信号を、前記人物位置選択工程で選択された人物位置に配置したものとしたときに観測して得られる第1の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第2の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定工程と、このフィルタ係数決定工程において決定されたフィルタ係数を用いて、前記音声入力工程において入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出工程と、を具備することを特徴とする音声収集方法。

【請求項10】 複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、複数のチャンネルを介して個々に音声を入力する音声入力工程と、前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、この人物位置選択工程において選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定工程と、

このフィルタ制約設定工程における制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力工程において入力される音声にフィルタ処理を施して音声を抽出する音声抽出工程と、を具備することを特徴とする音声収集方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は音声収集装置及び音声収集方法に関し、特に、音声認識装置やテレビ会議システムなどにおいて、雑音を取り除いて目的の音声を取り出す雑音抑圧技術に関する。

【0002】

【従来の技術】 音声認識装置やテレビ会議システムにおいては、高品質な音声の収録のため、指向性マイクロホンやマイクロホンアレイを使った雑音抑圧技術が提案されている。特に、テレビ会議システムの分野では、複数の会議参加者の中から発言者の音声と画像を自動的に得るため、ビデオカメラの画像を処理して得られた移動物体の位置に基づいて複数のマイクロホンの信号を処理する方法が例えば特開平5-227531号公報に開示されている。

【0003】 しかしながら、この方法ではマイクロホンアレイからの信号を、一つの目的の人物位置からの音声に対して位相が一致するようにした遅延和法により処理しているため、他の方向から到来した雑音に対する抑圧性能は十分でないという問題があった。

【0004】 一方、マイクロホンアレイの出力を処理して効果的に雑音を抑圧する技術としては、適応フィルタを指向性制御に用いた適応マイクロホンアレイ技術が従来より知られており、例えば文献（電子情報通信学会編 音響システムとデジタル処理 pp. 171-218）に詳述されている。適応マイクロホンアレイ処理では、雑音の到来方向を知る必要はないが、目的とする音波の到来方向は既知として処理するのが一般的である。音波の到来方向は、マイクロホンアレイからの信号を処理して推定することもできるが、発声中のみしか検出できないため処理の安定性に問題がある。

【0005】 これに対し、画像を処理して得られた人物位置を目的音の到来方向として用いる方法が知られており、この場合は、発声していないときにも位置が推定できるため安定であり、例えば文献（ICASSP '95 「Knowing Who to Listen to in Special Recognition Visually Guided Beamforming」 pp 848-851）に開示されている。

【0006】

【発明が解決しようとする課題】 しかしながら、上記文献をはじめとする従来の開示技術においては、画像の処理により人物位置が複数検出された場合に対する対処方法がないため、目的としない人物から発声があった場合はそれを除くような適応処理を行っていた。ところが、

この適応処理が完了するまでに妨害音が混入してしまったり、複数の話者が同時に発声した場合に、注目している一人の音声以外はクリアに入力できない、という問題があった。

【0007】本発明はこのような課題に着目してなされたものであり、その目的とするところは、複数の人物位置からの音声に対して、背景雑音を抑えてすべての音声を同時に抽出するかあるいは、特定の人物位置からの音声のみを抽出することが可能な音声収集装置及び音声収集方法を提供することにある。

【0008】

【課題を解決するための手段】上記の目的を達成するために、第1の発明に係る音声収集装置は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、複数のチャネルを介して個々に音声を入力する音声入力手段と、前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、任意に生成した音源信号を、前記人物位置選択手段によって選択された人物位置に配置したものとしたときに観測して得られる第1の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第2の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定手段と、このフィルタ係数決定手段によって決定されたフィルタ係数を用いて、前記音声入力手段によって入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出手段とを具備する。

【0009】また、第2の発明に係る音声収集装置は、第1の発明において、前記選択された人物位置のうち、前記特定の目的位置からの音声のみを高くするモードにおいて、複数の目的位置に対応して前記フィルタ係数決定手段及び音声抽出手段を複数個設け、複数の人物位置からの音声を分離して抽出する。

【0010】また、第3の発明に係る音声収集装置は、第1または第2の発明において、テスト発声データの入力と前記音声入力手段を介して入力される通常の音声入力の切り替えを指示する入力モード切り替え手段と、入力モードがテスト発声データ入力であるときに、取り込んだテスト発声データのレベルを求めるテスト発声レベル計算手段とをさらに具備する。

【0011】また、第4の発明に係る音声収集装置は、第1乃至第3の発明のいずれかにおいて、前記画像入力手段によって入力された画像から人物の発声動作に関する情報を位置別に検出する位置別発声動作情報検出手段

をさらに具備し、前記フィルタ係数決定手段は、検出した位置別の発声動作に関する情報と、入力された音声から求めた位置別到来パワーの少なくとも一方に基づいて、前記第1の信号である入力信号と前記第2の信号である希望応答信号とを生成する。

【0012】また、第5の発明に係る音声収集装置は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、複数のチャネルを介して個々に音声を入力する音声入力手段と、前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、この人物位置選択手段によって選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定手段と、このフィルタ制約設定手段の制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力手段によって入力される音声にフィルタ処理を施して音声を抽出する音声抽出手段とを具備する。

【0013】また、第6の発明に係る音声収集装置は、第5の発明において、前記フィルタ制約設定手段は、前記選択された人物位置の数が複数の場合に、この複数の人物位置の中の一つの位置を目的位置とし、該目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする第1の制約と、前記目的位置以外の人物位置からの音声に対しては、選択されなかった人物位置と比較して感度を低くする第2の制約をフィルタ処理の制約として設定し、前記音声抽出手段は、前記第1、第2の制約の基にフィルタ出力を最小化してフィルタ係数を決定する。

【0014】また、第7の発明に係る音声収集装置は、複数の人物を撮影して得られた画像を入力する画像入力手段と、この画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、複数のチャネルを介して個々に音声を入力する音声入力手段と、前記人物位置選択手段によって選択された少なくとも一つの人物位置の中の一つの位置を目的位置とし、この目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする制約を設定するフィルタ制約設定手段と、任意に作成した音源信号を、前記目的位置以外の人物位置に配置したものとしたときに観測される信号を生成する入力信号生成手段と、前記制約のもとで前記入力信号に基づき目的位置以外の人物位置からの音声に対して感度を低くするようにフィルタを決定するフィルタ決定手段と、このフィルタ決定手段によって求められたフィルタ

係数を用いて、前記音声入力手段によって入力された音声にフィルタ処理を施して音声を抽出する音声抽出手段とを具備する。

【0015】また、第8の発明に係る音声収集装置は、第7の発明において、前記フィルタ制約設定手段は、前記選択された人物位置の中から複数の目的位置を設定した場合に、この複数の目的位置の一つからの音声に対する感度を、選択されなかった人物位置と比較して高くする制約をフィルタ処理の制約として設定し、前記目的位置以外の人物位置に音源があるものとしたときに観測される入力信号に基づき、前記目的位置以外の人物位置からの音声に対しては感度を、選択されなかった人物位置と比較して低くするようにフィルタを設定するフィルタ設定手段と音声抽出手段とを、前記目的位置の変更に対応して複数個設け、複数の人物位置からの音声を分離して抽出する。

【0016】また、第9の発明に係る音声収集方法は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、複数のチャンネルを介して個々に音声を入力する音声入力工程と、前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、任意に生成した音源信号を、前記人物位置選択工程で選択された人物位置に配置したものとしたときに観測して得られる第1の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第2の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定工程と、このフィルタ係数決定工程において決定されたフィルタ係数を用いて、前記音声入力工程において入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出工程とを具備する。

【0017】また、第10の発明に係る音声収集方法は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、複数のチャンネルを介して個々に音声を入力する音声入力工程と、前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、この人物位置選択工程において選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定工程と、この

フィルタ制約設定工程における制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力工程において入力される音声にフィルタ処理を施して音声を抽出する音声抽出工程とを具備する。

【0018】

【発明の実施の形態】まず、本実施形態の概略を説明する。本実施形態では画像から人物位置を検出し、その人物位置に基づいて適応マイクロホンアレイにより雑音を抑圧して音声を抽出するシステムにおいて、特に、複数の人物位置が検出された場合に対処するため、以下の方法を用いる。

【0019】すなわち、第1の概略においては、雑音を抑圧するフィルタの決定に適応フィルタの処理方式の一つであるパイロット信号法を利用し、画像の処理により得られた複数の人物の位置に基づき適応フィルタの学習信号である入力信号と希望応答信号を人工的に生成し、複数の人物位置から到来する音波に対して同時に感度を一定に保つように制御を行なうことにより背景雑音を抑えてすべての話者の音声を同時に取り出すことを可能にするものである。

【0020】また、同じ構成により、目的の人物位置から到来する音波については感度が高く、他の人物位置から到来する音波については感度が低くなるように制御を行なうことにより、特定の人物のみの音声を取り出すことも可能にしている。パイロット信号法に関しては、上記文献（音響システムとデジタル処理）または文献（P. ROC. IEEE Vol. 55, No. 12 (1967)、B. Widrow: 「Adaptive Antenna systems」）に詳述されている。

【0021】また、第2の概略においては、適応フィルタによる雑音抑圧のフィルタ係数の学習の際、人物位置から到来する音波のパワーを推定し、このパワーに基づいて適応フィルタの入力信号の振幅と適応フィルタの収束速度を決定しているため、人工的に生成する信号を実際の環境に忠実に合わせることができ、精度良く雑音抑圧のフィルタを学習することができる。

【0022】また、第3の概略においては、適応フィルタによる雑音抑圧のフィルタ係数の学習の際、人物位置から到来する音波のパワーに加え、人物の画像から発声動作を表す情報を抽出しており、これらの位置ごとのパワーと位置ごとの発声動作に関する情報から、適応フィルタの入力信号の振幅と適応フィルタの収束速度を決定しているため、背景雑音が大きく、到来パワーの推定精度が低い場合でも精度良く雑音抑圧のフィルタを学習することができる。

【0023】また、第4の概略においては、テスト発声を収集するための入力モード切り替え手段を具備することにより、テスト発声データを入力し、その入力レベルからパイロット信号法によって適応フィルタ処理を行う際の学習信号の振幅を正確に決定して適応フィルタの学

習を行うことにより、高精度な雑音抑圧を可能としている。

【0024】また、第5の概略においては、目的の人物から到来する音に対しては感度を一定に保ち、他の人物から到来する音に対しては感度を低くするように適応フィルタの学習信号を生成してフィルタ係数を学習し、人物ごとにこのようなフィルタを用意することにより、複数の人物が発声した場合に、各人物ごとの発声を別個に取り出すことを可能にしている。

【0025】また、第6の概略においては、パイロット信号法による適応フィルタ処理の代わりに、拘束条件付き適応フィルタ処理を用い、画像の処理により得られた複数の人物の位置に対してマイクロホンアレイの感度を一定に保つという拘束条件のもとで適応フィルタの係数を決定することにより、背景雑音を抑えてすべての話者の音声を同時に取り出すことを可能にしている。この方式では、パイロット信号法で必要だった発声区間に応じた適応の制御が不要であり、より少ない構成要素で実現できる。

【0026】同概略においては、目的の人物から到来する音に対しては感度を一定に保ち、他の人物から到来する音に対しては感度を低くするという制約のもとで雑音抑圧のフィルタを決定することにより、特定の人物の音声だけを取り出すことも可能である。

【0027】また、第7の概略においては、第6の概略において用いた拘束条件付き適応フィルタ処理を用い、目的の人物から到来する音に対しては感度を一定に保ち、他の人物から到来する音に対しては感度を低くするという制約のもとで雑音抑圧のフィルタの係数を決定し、人物ごとにこのようなフィルタを用意することにより、複数の人物が発声した場合に、各人物ごとの発声を別個に取り出すことを可能にしている。

【0028】また、第8の概略においては、第6、第7の概略で用いた拘束条件付き適応フィルタ処理と第1から第5の概略で用いたパイロット信号法を組み合わせ、目的の人物から到来する音に対しては感度を一定に保ち、制約のもとで雑音抑圧のフィルタの係数を決定し、他の人物位置から到来する音に対しては感度が低くなるように学習信号を生成し、この学習信号により拘束条件付き適応フィルタによりフィルタ係数を決定することにより、拘束条件が多くなることによる雑音抑圧性能の低下を抑え、また学習信号生成の演算量を減らして同等の性能を実現している。

【0029】また、第9の概略においては、第8の概略において雑音抑圧のフィルタを人物ごとに複数用意することにより、複数の人物が発声した場合に、各人物ごとの発声を別個に取り出すことを可能にしている。

【0030】以下に図面を参照して上記した実施形態を詳細に説明する。

【0031】まず、図1を参照しながら、第1実施形態

について説明する。本実施形態は、画像を処理して検出した人物位置に基づいて適応フィルタの学習信号を生成し、学習したフィルタにより雑音抑圧処理を行うものである。本実施形態では、複数の人物位置を対象として適応フィルタ学習の制御を行えるようにしているため、従来1個の対象にしか考えられてこなかった雑音抑圧処理を、複数の対象に対して行え、会話や複数話者の同時発声の場合の音声入力を自動的に安定して高対雑音比で行うことが可能である。

【0032】図1において、1はビデオカメラなどから画像を入力する画像入力部、2は入力した画像を処理して人物の位置を検出する人物位置検出部、3は複数のマイクロホンからの音声を並列に入力する音声入力部、4は画像処理により検出された人物位置に基づいて複数のマイクロホンからの入力音声から雑音を抑圧して音声を取り出す雑音抑圧部であり、その内部構成は、人物位置検出部2により検出された人物位置の中から処理対象とする人物位置を選択する人物位置選択部4-1と、選択された人物位置に基づいて適応フィルタの学習を行う環境適応部（フィルタ係数決定手段）4-2と、決定されたフィルタ係数により雑音抑圧処理を行う雑音除去部4-3とからなる。

【0033】この構成において、画像入力部1より入力した画像を人物位置検出部2に送り、人物位置検出部2において人物の方向または位置を画像から検出する。検出した人物位置の中から処理対象とする人物位置を人物位置選択部4-1において選択し、環境適応部4-2において、前記選択された人物位置に基づいて適応フィルタの学習信号を生成して適応フィルタの係数を決定し、雑音除去部4-3において決定したフィルタ係数を用いて入力音声にフィルタ処理を行って雑音を抑圧する。

【0034】以下に上記した処理をさらに詳細に説明する。まず、画像からの人物位置の検出について説明する。画像からの人物位置の検出は、テンプレートマッチングに基づいた顔領域の抽出処理などにより行なうことができ、テンプレートマッチングについては例えば「画像解析ハンドブック」（東京大学出版会）に詳述されている。さらに、テンプレートマッチングを用いて画像中の物体の動きを追跡することができ、例えば情報処理学会技術報告CV76-7, pp. 49-56 (1992), 小杉他：「シーンの中の顔の探索と認識」に詳述されている。これらの開示技術により、同じ人物の座標を時間ごとに特定し追跡できることが知られている。なお、これらの技術では、人物の顔を含む小領域の画像を扱うため、人物位置の検出とともにこれらの画像も得ることができる。

【0035】画像による人物位置の検出では、一般に、画像入力に1個のビデオカメラを用いた場合、人物のカメラに対する方向は十分な精度で特定することができるが、カメラと人物の間の距離方向の測定は誤差が大き

い。それでも、人物の大きさの比較により、距離方向に関しておおまかな相対関係を得ることは可能である。ステレオカメラを用いた場合は、距離に関して高精度に測定できるが、本実施形態では、人物の方向とおおまかな距離関係がわかれば十分であるため、カメラ1個による人物位置の抽出手法を用いればよく、ステレオカメラは必ずしも必要ない。もちろんステレオカメラあるいは多数のカメラを用いても差し支えない。人物位置の検出方法は本実施形態の本質に関わりがないので詳しい説明は省略するが、現状で利用可能な技術であることは言うまでもない。

【0036】ビデオカメラとマイクロホンを組み合わせて処理を行う際、ビデオカメラとマイクロホンの位置の設定は種々考えられるが、ビデオカメラが1個の場合、例えば図2(a)に示すように設定する。ビデオカメラ5とマイクロホンアレイ6は人物から見て同じ方向にあるように設置し、マイクロホンアレイ6の処理とビデオカメラ5とで共通の方位座標を使うようにするのが望ましい。

【0037】なお、図2(b)に示すように、ビデオカメラ5を複数使う場合は、人物位置が3次元座標として得られるので、先のようにカメラ位置とマイクロホン位置を利用者から見て同じ方向に設定する必要はなく、マイクロホンアレイ処理の際、ビデオカメラ5から得られた人物座標をマイクロホンアレイ6からみた角度に変換して用いることもできる。

【0038】以上の処理により人物位置が得られた後、雑音抑圧部4においては、人物位置選択部4-1により人物位置の中から処理対象とする人物位置を予め決めた数だけ選択し、該選択した人物位置に基づき、環境適応部4-2により適応フィルタの学習信号を生成して適応フィルタに入力し、フィルタ係数を決定する。そして、決定したフィルタ係数を用い、雑音除去部4-3で複数のマイクロホンからの入力音声に上記のフィルタによるフィルタ処理を行って出力音声を取り出す。

【0039】複数のマイクロホンから入力した音声を処理して雑音を抑圧するための適応フィルタとしては種々のものが知られており、例えば文献(Haykin著: Adaptive Filter Theory)に詳述されているが、本実施形態では、複数の任意の方向あるいは位置から到来する音に対するアレイの応答を比較的簡単に設定できるパイロット信号法を用いている。

【0040】以下に、雑音抑圧部4の詳細を説明する。まず、雑音抑圧部4では、人物位置選択部4-1において複数の人物位置の中から、音声の抽出処理を行う対象の人物位置を選択する。この選択においては、選択する人物位置の最大数をN、例えば $N=3$ とし、人物位置検出部2で特定された人物位置の数がNより大きい場合に、特定された人物位置の中からN個の位置を選択し、小さい場合はすべてを選択する。選択の方法に関して

は、例えば、カメラと人物の距離を基準とし、この距離が小さい順にN個の位置を用いるようにしてもよいし、カメラの中心方向と人物方向の角度差を基準とし、この角度差が小さい順にN個の位置を用いてもよい。

【0041】また、上記2つの選択基準を組み合わせた値を基準としてもよい。すでに述べたように、使用するカメラが1個の場合で、カメラと人物の間の距離の値を得ることが困難な場合は、人物の大きさまたは人物の顔の大きさを人物位置とカメラとの距離の目安として使うことが可能である。

【0042】例えば、図3に示すような画像データから図4に示すような人物位置の方向(X, Y)と顔の大きさ(A)、および人物方向から計算されるカメラ中心線方向と人物方向との角度差(B)が得られた場合、顔部分の面積が大きいほどカメラに近いとしてこの面積が大きい順に人物の番号6, 4, 3の3人を選択してもよいし、カメラ中心線方向と人物方向との角度差が小さい順に人物番号4, 2, 5を選択してもよいし、上記A, Bを組み合わせた値、例えば A/B の値の大きい順に人物番号4, 3, 6を選択してもよい。

【0043】次に、パイロット信号法による適応フィルタ処理を行うため、環境適応部4-2は、図5に示すような構成としている。図5において、4-2aは入力信号生成部、4-2dは希望応答生成部、4-2eは適応処理部、4-2cは学習信号レベル計算部、4-2bは音源信号生成部である。

【0044】この構成において、まず、音源信号生成部4-2bにより、選択された人物位置ごとに音源があるものと仮定してその発生信号を生成し、学習信号レベル計算部4-2cにより、入力音声に基づいて入力信号生成の際の音源信号のレベルを決定する。次に、求められた学習信号レベルと音源信号とから、入力信号生成部4-2aにより、選択された人物位置に基づき、適応フィルタの入力信号を生成すると同時に、学習信号レベルと音源信号とから、希望応答生成部4-2dにより適応フィルタの希望応答を生成し、生成した入力信号と希望応答を適応処理部4-2eに入力し、適応フィルタの適応処理を行う。適応フィルタの処理方式は、よく知られたLMSでもまた、RLSでもよく、文献(ヘイキン著: 適応フィルタ入門)に詳述されている。ここでは、LMS適応フィルタにより説明する。

【0045】適応フィルタの処理は、複数チャネルの入力各々に対し、図6に示すような遅延線タップ付きフィルタから構成されるユニバーサル型フィルタを用いて行うようにしている。図6において、フィルタのタップ数をJ、i番目のマイクロホンのフィルタ係数を w_{ij} 、

($1 \leq i \leq N$, $1 \leq j \leq J$)としており、Jは例えば200を用いる。この構成において、i番目のマイクロホンの波形を $x_i(n)$ とし、時刻nにおいてJサンプル過去から時刻nまでの各マイクロホンの波形サンプルの系

列 $x_i = (x_{i(n-J+1)}, x_{i(n-J+2)}, \dots, x_{i(n-1)}, x_{i(n)})$ を全マイクロホンについて並べ、
 $X = (x_1, x_2, \dots, x_N)^T$ (1)

と、ベクトルで表す。また、 i 番目のマイクロホンのフイルタ係数 w_{ij} を並べてベクトルで表して

$$w_i = (w_{i1}, w_{i2}, \dots, w_{iJ}) \quad (2)$$

とし、さらに全マイクロホンについて並べて

$$W = (w_1, w_2, \dots, w_J)^T \quad (3)$$

と表す。式 (1)、(3) から、フィルタの出力は、

$$Y = W^H X \quad (4)$$

と表される。ここでフィルタ係数 W の要素は複素数と

し、 H はベクトルの複素共役転置を表すものとする。 X は一般にスナップショットと呼ばれる。

【0046】 LMS 適応フィルタ (Normalized LM ☆

$$W_j = W_{j-1} - a * e * X / 2p \quad (5)$$

ここで、 W_j は j 回の更新後のフィルタ係数、 e は誤差信号 $e = d - W^H X$ 、 d は希望応答、 p は希望応答のパワー、 a はステップサイズであり、 $0 < a < 1$ 、 0 の範囲で実験的に決められるが、例えば 0.1 などが用いられる。

【0048】 上記のフィルタ更新に用いる入力信号 X と希望応答 d は、人物位置に基づき入力信号生成部 4-2 a と希望応答生成部 4-2 d で音源信号から各々生成する。これらの信号は人工的に生成するものであり、信号の内容によって雑音抑圧の仕方を制御することができる。例えば、選択された人物位置すべてから到来する音波に対して感度を高くする (A) ことや、選択された人物位置のうち、ある人物位置からの音波に対しては感度を高くするが、それ以外に対しては抑圧する (B) などのように制御できる。

【0049】 今後、上記 2 つの抑圧処理の仕方を、抑圧処理のモード (A)、(B) と呼ぶことにする。特にモード (B) は、妨害音の発生する可能性の大きい方向に対して事前に感度を低くする方法であり、従来の適応マイクロホン処理で行われていたように、妨害音が発生してからその環境に適応して抑圧する手法よりも大幅に高品質な音声入力が行える。抑圧処理モードの設定は、初期設定の際に環境適応部 4-2 において設定するようにする。

☆

$$\tau_i(\theta) = ((x_i - x_1)^2 + (y_i - y_1)^2)^{1/2} \times \cos(\theta - \tan^{-1}((y_i - y_1) / (x_i - x_1))) \quad (6)$$

$$\text{振幅は } a_1 = a_2 = \dots = a_N = 1 \quad (7)$$

とおくことができ、点音源を仮定した場合、図 7 (b) ◇ のように仮想音源位置 θ を (x_s, y_s) とおくと

$$\begin{aligned} \tau_i &= ((x_i - x_s)^2 + (y_i - y_s)^2)^{1/2} \\ &\quad - ((x_1 - x_s)^2 + (y_1 - y_s)^2)^{1/2} / c \quad (8) \\ \text{振幅は } a_i &= ((x_i - x_s)^2 + (y_i - y_s)^2)^{1/2} \\ &\quad / ((x_1 - x_s)^2 + (y_1 - y_s)^2)^{1/2} \quad (9) \end{aligned}$$

となる。ただし、 c は音速である。なお、ここでは簡略化のため 2 次元平面上で説明したが、3 次元空間への拡張は容易である。

【0053】 上のようにして求めた遅延時間 τ_i を用 *

$$x_i(n) = S_k(n - \tau_i') \quad (10)$$

☆ S) による適応処理部 4-2 e では、上記のフィルタ構造において次式に従ってフィルタ係数を更新し、フィルタ係数の学習を行う。

【0047】

☆ 【0050】 フィルタ更新に用いる入力信号 X と希望応答 d の生成の前段階として、まず、音源信号生成部 4-2 b において、人物位置の数の信号系列である音源信号を発生する。発生した音源信号の内容は人工的なものでよく、例えば、ランダム雑音でもよい。このとき、ランダム雑音は人物位置ごとに無相関となるようにする。また、人物位置ごとに独立な乱数系列から生成するようにする。また、周波数特性は、例えば平均的な音声のスペクトルの傾きと同じになるようにフィルタをかけてもよい。

【0051】 次に、入力信号生成部 4-2 a では、生成した音源信号が空中を伝播してマイクロホン位置に到達すると仮定したときのマイクロホンで観測される信号を計算する。マイクロホン位置で観測される信号は、音源信号の伝搬時間差と伝搬に伴う振幅変化から計算できる。

【0052】 例えば、マイクロホンと人物位置が図 7 のような設定であるとして、図 7 を参照して次のように行う。図 7 (a) のように、1 番目のマイクロホンの座標を (x_1, y_1) 、 i 番目のマイクロホンの座標を (x_i, y_i) とすると、平面波を仮定した場合、 θ 方向から音波が入射する際の i 番目のマイクロホンと 1 番目のマイクロホンに入射する音波の伝搬時間差 τ_i は、

とできる。ここで、 τ_i' は τ_i を四捨五入した値である。また、信号の遅延をもっと精度よく行うため、四捨五入する代わりに、上記した音響システムとデジタル処理 (pp. 215) に述べられているようにデジタルフィルタを畳み込んでよいし、フーリエ変換により周波数領域に変換して位相回転により遅延を与えた後、逆フーリエ変換してもよい。

【0054】次に、学習信号レベル計算部4-2cにおいて音源信号のレベルを決め、以上のようにして求めたマイクロホン位置での音源信号の観測値が、決定したレ

$$A_k = (P_N * 10^{v/10})^{1/2}$$

により計算する。ここで、 A_k は音源信号の振幅、 P_N は背景雑音のパワーである。

【0056】次に、希望応答生成部4-2dでは、前記の2つの抑圧処理モード(A)、(B)に応じて、別の方法で希望応答を生成する。選択された人物位置すべてから到来する音波を収集する場合(A)は、選択されたすべての人物位置から音波が到来すると仮定したときのマイクロホン位置での観測信号を希望応答として出力するようにする。この場合、例えば、1番目のマイクロホン位置で観測される信号を希望応答として使うようにする。ただし、マイクロホン位置で観測される信号よりも遅延させたものとするようにする。遅延の大きさは、例えばタップ数の半分とする。

【0057】選択された人物位置のうち、ある人物位置からの音波について抑圧したい場合(B)では、人物位置から音波が到来すると仮定したときのマイクロホン位置での観測信号作成の際に、その人物位置からの音波に相当する信号は加算しないようにする。例えば、選択した人物位置が3個で、入力したい人物位置がその中の1個の場合は、入力したい1個の人物位置からの到来だけを仮定してマイクロホンで観測される信号を希望応答とする。

$$P_N = \gamma * P_N + (1 - \gamma) P_N'$$

ここで、 P_N' は、それまでに求まっていた背景雑音パワー、 γ は忘却係数であり、例えば、 $\gamma = 0.1$ である。

【0061】次に、学習信号レベル計算部4-2cにおいて、音源信号が伝播してマイクロホン位置で観測されると仮定したときの信号を計算し、式(11)により音★

$$x_i(n) = r_i(n) + \sum A_k S_k(n - \tau_k') \quad (13)$$

により計算する(ステップS4)。次に、希望応答生成部4-2dにおいて、音源信号と音源信号の振幅から★

$$d(n) = \sum A_k S_k(n - \tau_k' - n_0) \quad (14)$$

となる。ここで、 n_0 は適当な遅延、例えば、 $n_0 = 10$ である。ただし、抑圧処理のモードがAの場合、 k はすべての人物位置について変化させ、モードがBの場合

*算のため、学習信号レベル計算部4-2cにおいては、入力音声の背景雑音パワーの音声区間の平均値を計算して保持するようにする。入力音声の背景雑音パワーは、例えば複数ある中の1番目のマイクロホンのパワーを逐次計算して音声区間を検出し、音声区間として検出されなかった区間の平均パワーを求めるようにする。パワーに基づいた音声区間検出はよく知られているように、例えば文献(新美著：音声認識)に詳述されている。

【0055】このようにして求めた背景雑音パワーに対して一定値 v 、例えば、 $v = 7$ dB高い値を音源信号レベルとし、このパワーの平方根の値を音源信号の振幅とするようにする。すなわち、

$$(11)$$

※【0058】以上に述べた環境適応部4-2を含む雑音抑圧部4における音声など連続信号の処理は例えば、1chあたり1024点を1ブロックとし、ブロック単位で行うようにする。すなわち、音声入力部3における音声データの読み込み、環境適応部4-2における音源信号と学習信号の生成、適応フィルタ処理、雑音除去処理などは、すべて1chあたり1024点を1ブロックとしてブロック単位で行うものとする。

【0059】ここで、以上に述べた環境適応部4-2の処理の流れについて図8を参照しながら説明する。

【0060】まず、環境適応部4-2の音源信号生成部4-2bにおいて、選択人物位置の数の系列の音源信号を生成する(ステップS1)。音源信号は音源ごとに無相関な系列とし、分散は1に正規化しておくようにする。次に、学習信号レベル計算部4-2cにおいて、複数チャンネルで入力した入力音声の中から、例えば1ch目の信号のパワーを、例えば波形128点の小セグメントごとに計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーを求め、これを背景雑音パワー P_N とする(ステップS2)。背景雑音パワーは、それまでに求まっていた値との間で平均化してもよく、その場合、次式により平均化する

$$(12)$$

★源信号の振幅 A_k を計算する(ステップS3)。次に、入力信号生成部4-2aにおいて、実際の入力音声と加算して適応フィルタの入力信号を生成する。すなわち、 i ch目の入力音声を $r_i(n)$ とすると、 i ch目の適応フィルタの入力信号 $x_i(n)$ は、

☆希望応答を生成する。式で表すと、

合、 k は感度を高く設定する人物位置について変化させる。式(12)(13)のように、音源信号の遅延をサンプリング周期で四捨五入した値 τ_k' により遅延を与

えるかわりに、もっと精度よく遅延させることも可能であることはすでに述べた（ステップS5）。

【0062】次に、生成した入力信号と希望応答を適応フィルタに入力し、フィルタ係数を得る（ステップS6）。得られたフィルタは、雑音除去部4-3に送り、入力音声进行处理して音声を抽出する。雑音除去部4-3におけるフィルタ処理は、式（4）に従って行う。

【0063】次に、図9を参照しながら本実施形態全体の処理の流れを説明する。

【0064】まず、初期設定を行い、選択する人物位置の数Nと雑音抑圧処理のモードAかBかを設定する（ステップS31）。

【0065】画像の処理の方では、画像データをビデオカメラ5から、例えば毎秒5フレームで取り込み（ステップS32）、フレームごとに人物位置を特定して出力し（ステップS33）、このステップS32とS33を繰り返す。画像から人物位置を特定する画像の処理は、音声の雑音抑圧処理とは独立に、並列に処理するようにする。

【0066】音声処理の方では、まず、音声データを、例えばサンプリング12kHzでAD変換し、1チャンネルあたり、例えば1024サンプルを1ブロックとして1ブロック分のデータを取り込む（ステップS34）。次に、人物位置が特定されているか否かを判定し（ステップS35）、人物位置が特定されていない場合は、何もせずにこのステップS34とS35を繰り返し、特定された場合は次のステップS36に進む。人物位置は、処理開始直後で画像処理結果が出ていない場合や人物がない場合に特定されない。位置画像に関する処理と音声に関する処理とは独立しているため、例えば、一つの計算機上で全処理を行う場合、よく知られているように、ソケットを用いたプロセス間通信やシェアドメモリ、あるいはファイルを通じて人物位置のデータのやり取りを行うことができる。

【0067】次に、ステップS36では、人物位置選択部4-1において、処理対象とする人物位置を選択する。次に、環境適応部4-2において、人物位置選択部4-1で選択された人物位置または方向と距離を用いて適応フィルタの学習信号を生成し、フィルタ係数を更新する（ステップS37）。学習信号の長さは、取り込んだ音声データの長さと同じく1chあたり1024点にする。

【0068】次に、ステップS37で更新されたフィルタ係数を雑音除去部4-3にコピーし、このフィルタと入力音声との畳み込み演算を行って音声を出力する（ステップS38）。以上のステップS31からS32の処理とS33からS38までの処理を並列に繰り返す。

【0069】以上に述べた処理により、画像処理により特定された複数の人物位置各々から到来する音声の感度を設定できるように雑音抑圧処理を行うフィルタの係数

を学習しているため、複数の人物が同時に発声した場合に、その人物すべての音声を背景雑音を抑圧して取り出したり、一人だけの人物の音声のみを他の人物の音声を抑圧して取り出すことが可能となる。

【0070】また、逐次人物位置を特定し、その人物位置に応じてフィルタ処理の学習信号を生成しているため、複数の人物が各々動いた場合でも追従して雑音抑圧処理を行うことが可能である。

【0071】以下に本発明の第2実施形態を説明する。第2実施形態では、音声処理の対象として選択された人物各々からの発声音を検出し、この検出情報に基づいて学習信号の生成を制御することによって、学習を高精度に行う。

【0072】第1実施形態で述べたパイロット信号法による適応フィルタの学習では、人物が発声中であるか否かに関わらず学習を行っていたが、感度を高くして入力したい人物が発声している間は適応を止めることにより、また、抑圧したい人物の発声中は、その人物方向からの到来を仮定した音源信号を使わずに入力信号と希望信号を生成することにより、より環境に適応した高精度な雑音抑圧の学習が行える。このため、本実施形態では、人物位置ごとに発声中であるかどうかの目安となる位置別の到来パワーを推定する位置別到来パワー推定部4-4をさらに具備しており、これを含んだ全体構成を図10に示す。図10において、4-1は人物位置選択部、4-2は環境適応部、4-3は雑音除去部である。

【0073】また、推定した位置別到来パワーに基づいてフィルタ学習の制御を行うため、環境適応部4-2は図11のような構成を具備している。図11において、4-2aは適応フィルタの入力信号生成部、4-2dは適応フィルタの希望応答生成部、4-2eは適応フィルタによる適応処理部、4-2bは入力信号と希望応答生成の際に用いる人工的な波形である音源信号を発生する音源信号生成部、4-2cは入力信号と希望応答の生成の際、人物位置ごとの音源信号の振幅を位置別到来パワーに基づいて決定する学習信号レベル計算部、4-2fは位置別到来パワーからフィルタ学習の際の適応速度を制御するパラメータを決定する適応制御信号制御部である。

【0074】位置別到来パワー推定部4-4では、マイクロホンアレイ6に入力した音声から、人物位置ごとの到来パワーを求める。マイクロホンアレイ6によって位置あるいは方向ごとの到来パワーを計算する方法としては、文献（音響システムとデジタル処理）に詳述されているように、遅延和法、最小分散法、MUSIC法など種々の方法があるが、ここでは、少ない計算量で実現可能な遅延和法による方法を説明する他の方法も計算量が多くなるだけで適用可能であることは言うまでもない。

【0075】上述の文献にも詳述されているように、遅延和法は、各マイクロホンからの信号を対象とする方向

または位置から到来する音波の位相が揃うように遅延させてから和をとるものである。図2に示すようなマイクロホンと到来位置の関係の場合、 i 番目のマイクロホンと1番目のマイクロホンに入射する音波の伝搬時間差 τ_i は、平面波が入射する場合は式(6)により、球面波が入射する場合は式(8)により計算できる。このと *

$$p = |\sum x_i(n - \tau_i)|^2 / M \quad (15)$$

であり、この値は、対象とする方向または位置から音波が到来している場合には音源のパワーに比例することが知られているので、式(15)により、各人物位置から到来するパワーが推定できる。なお、球面波の場合は音源とマイクロホンとの距離により補正係数が必要になるが、容易に補正できる。詳細は多数センサによる音源波形推定に関する文献(日本音響学会誌、47、4、pp 268-273、1991)に述べられている。

【0076】次に、学習信号レベル計算部4-2cについて説明する。ここでは、求められた位置別到来パワーから、適応フィルタの入力信号と希望応答の生成の際の、各々の人物位置の音源信号の振幅を決定する。このため、学習信号レベル計算部4-2cでは、入力音声の背景雑音パワーと位置別到来パワーの音声区間の平均値を計算して保持するようにする。入力音声の背景雑音レベルは、複数のマイクロホン中の、例えば1番目のマイクロホンのパワーを逐次計算して音声区間を検出し、音声区間として検出されなかった区間の平均パワーを求め※

$$A_k = (P_N * 10^{v/10})^{1/2}$$

(位置別到来パワーが背景雑音+ v dBより小さいと ★ ★き)

$$A_k = A_{k0} = (P_k)^{1/2}$$

(位置別到来パワーが背景雑音+7 dBより大きいとき)のように計算する。ここで、 P_N は背景雑音のパワー、 P_k は k 番目の位置の位置別到来パワーである。 ☆

$$A_k = A_{k0} * (P_N / P_k)^{1/2}$$

によって計算する。以上のようにして求めた音源振幅と入力音声とを加算し、適応フィルタの入力信号を生成する。

【0080】例えば、選択された人物位置が a と b の2個であり、 a が音声入力の対象とする目的の人物位置であり感度を高く設定する位置、 b が感度を低く設定する妨害音の位置であるとする。図12の(1)、(2)に示すように a 、 b の位置ごとの到来パワーが推定された場合、1、2で示した区間では、(3)に示すように入力信号中の a に関する成分を大きくし、3で示した区間では(4)に示すように b に関する成分を小さくする。また、希望応答は、入力信号中の a に関する成分と同じとし、 b に関する成分はすべて0とするか加算しないようにする。 ◆

$$a = C / (\alpha_B P_k / P_N + 1) \quad (\text{抑圧モードBのとき}) \quad (19)$$

$$a = C / (\alpha_A \sum (P_k / P_N) / M + 1) \quad (20)$$

(抑圧モードAのとき)

ここで、 P_N は背景雑音のパワー、 P_k は k 番目の位置

*き、 i 番目のマイクロホンの波形を $x_i(n)$ とし、時刻 n において J サンプル過去から時刻 n までの各マイクロホンの波形サンプルの系列 $x_i = (x_i(n-J+1), x_i(n-J+2), \dots, x_i(n-1), x_i(n))$ を τ_i 遅延させた場合の全マイクロホンについての平均パワーは、

※るようにする。パワーに基づいた音声区間検出はよく知られているように、例えば文献(新美著:音声認識)に詳述されている。また、位置別到来パワーに関しても同様に位置ごとに音声区間の検出を行い、こちらは音声区間の平均パワーを求めるようにする。

【0077】このようにして求めた位置別到来パワーと背景雑音パワーから、音源信号の振幅を計算する。このとき、人物位置が、感度を高く設定した位置であるか、低く設定した位置であるかに応じて振幅の計算方法を変えるようにする。

【0078】感度を高くするように設定した位置の場合、上記のように求めた位置別到来パワーの位置ごとの平均値の平方根の値を音源信号の振幅とするようにする。なお、発声がない場合は、位置別到来パワーは小さい値となるため、位置別到来パワーが背景雑音に対してある値 v 、例えば、 $v = 7$ dBを上回る時だけ位置別到来パワーの平方根の値に設定するようにする。すなわち、 A_k を k 番目の位置の音源信号の振幅とすると、

$$(16)$$

$$(17)$$

☆【0079】感度を低くするように設定した位置の場合、位置別到来パワーが大きいほど小さい振幅となるようにする。例えば、 k 番目の位置の音源信号の振幅を、

$$(18)$$

◆【0081】次に、位置別到来パワーに基づいた適応フィルタの適応速度の制御について説明する。適応フィルタの学習は、よく知られているように、式(5)のステップSサイズの値 a により制御できる。ここでは、音声入力の対象となる位置からの到来パワーが大きい場合は、入力信号中の抽出すべき信号があるにも関わらず希望応答の中にその信号がないため、抑圧の対象となってしまう。そこで、この到来パワーの値が大きいときは適応を遅くあるいは停止し、小さいときは適応を早くするように適応速度を制御する。

【0082】このため、例えば、式(5)で固定していたステップサイズ(式(5)の a)の値を次式により逐次計算して可変とするようにする。

【0083】

の位置別到来パワー、 C 、 α_A 、 α_B は定数、例えば、 C

= 2. 0, $\alpha_A = \alpha_B = 1$ である。ステップサイズの計算式として挙げた上式は一例であり、他の方法も使用可能である。

【0084】ここで、図13を参照しながら第2実施形態の環境適応部全体の処理の流れを説明する。

【0085】まず、環境適応部4-2の音源信号生成部4-2bにおいて、選択人物位置の数の系列の音源信号を生成する(ステップS11)。次に、学習信号レベル計算部4-2cにおいて、複数チャンネルで入力した入力音声のパワーを計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーから背景雑音パワー P_N を求める(ステップS12)。このとき、式(12)により平均化してもよい。次に、複数チャンネルの入力音声から式(15)により位置別到来パワーを計算する(ステップS13)。

【0086】次に、学習信号レベル計算部4-2cにおいて、式(16)から式(18)により音源信号の振幅 A_k を計算する(ステップS14)。次に、入力信号生成部4-2aにおいて、式(13)により、実際の入力音声と加算して適応フィルタの入力信号を生成する(ステップS15)。次に、希望応答生成部4-2dにおいて、式(14)により、音源信号と音源信号の振幅から希望応答を生成する(ステップS16)。次に、適応制御信号生成部4-2fにおいて、式(19)または(20)により、背景雑音パワーと位置別到来パワーからステップサイズの系列である適応制御信号を生成する(ステップS17)。

【0087】次に、生成した入力信号と希望応答と適応制御信号を適応フィルタに入力し、フィルタ係数を得る(ステップS18)。得られたフィルタは、雑音除去部4-3に送り、入力音声とフィルタを畳み込んで音声抽出する。雑音除去部4-3におけるフィルタ処理は、式(4)に従って行う。

【0088】第2実施形態の全体の処理の流れは第1実施形態と同じであるので改めて述べない。

【0089】以上に述べたように、画像処理により特定された複数の人物位置各々から到来する音声に対し、適応フィルタにより感度を設定して雑音抑圧を行う際、人物位置からの到来音のパワーに応じて適応フィルタの適応処理を制御しているため、実環境に応じた高精度な適応が行え、雑音抑圧性能を大幅に高くしながら、複数の人物が同時に発声した場合に、その人物すべての音声を背景雑音を抑圧して取り出したり、一人だけの人物の音声のみを他の人物の音声を抑圧して取り出すことが可能となる。

【0090】また、画像から逐次人物位置を特定し、その人物位置に応じてフィルタ処理の学習信号を生成して*

$$K(i) = \sum_x \sum_y |G(i, x, y) - G(i-1, x, y)|^2$$

* いるため、複数の人物が各々動く場合でも追隨して雑音抑圧処理を行うことができる。

【0091】以下に、音声パワー検出と画像からの発声動作検出を行なう第3実施形態について説明する。第3実施形態は、第2実施形態において行っていた人物位置ごとの到来パワー推定に加え、画像データに基づいた発声動作の検出を行い、これら2つの情報に基づいて学習信号の生成と適応速度の制御を行うことにより、音の環境をより正確に反映して適応フィルタの学習を行うようにする。本実施形態では画像から発声動作を検出しているため、高雑音下でも人物が発声中かどうかを精度よく検出でき、高精度な適応フィルタの学習の制御が行える。

【0092】画像に基づいた発声動作の検出と位置別到来パワーに基づいて適応フィルタの学習を制御するため、本実施形態の雑音抑圧部は、第2実施形態の雑音抑圧部の構成にさらに画像から発声動作に関する情報を検出する発声動作情報検出部を追加し、図14のような構成としている。

【0093】図14において、1はビデオカメラなどから画像を入力する画像入力部、2は入力した画像を処理して人物の位置を特定する人物位置検出部、3は複数のマイクロホンからの音声を並列に入力する音声入力部、4は画像処理により検出された人物位置に基づいて複数のマイクロホンからの入力音声から雑音を抑圧して音声を取り出す雑音抑圧部である。

【0094】雑音抑圧部4は、人物位置検出部2により特定された人物位置の中から処理対象とする人物位置を選択する人物位置選択部4-1と、選択された人物位置に基づいて適応フィルタの学習を行う環境適応部4-2と、決定されたフィルタ係数により雑音抑圧処理を行う雑音除去部4-3と、人物位置ごとに到来パワーを検出する位置別到来パワー推定部4-4と、人物位置ごとに画像から発声動作に関する情報を検出する位置別発声動作情報検出部4-5とからなる。

【0095】画像による音声区間の検出は、口元の画像の時間変化から行う方法が知られており、簡単には、口元画像全体の輝度変化を時刻ごとに計算し、その変化が大きい時刻を発声中であるとして検出できる。ここでは、正確な口元画像の代わりに人物位置検出部2で特定した人物の顔を含む画像において、例えばその下半分についての画面全体にわたる輝度の時間変化を求めて発声動作の目安とするようにする。画像データのフレームの番号を i 、縦横位置 x, y における人物位置の顔を含む顔周辺画像データを $G(i, x, y)$ とすると、フレーム i と $i-1$ の間の輝度変化は、

$$(20-1)$$

と計算でき、この $K(i)$ の値を発声動作があるか否かの目安とする。人物の顔周辺画像は、人物位置検出部2か

ら、特定した位置とともに入力するようにする。人物位置特定処理では、全体画像中から顔の部分の画像を切り出す処理を含むのが一般的であるため、顔周辺画像は容易に取り出せる。画像からの発声動作の検出方法は輝度変化の計算に限るものではなく、他の方法も使用可能である。以降上の輝度変化 $K(i)$ を含め、画像から抽出した発声動作の目安となる情報を便宜上発声動作情報と呼ぶことにする。

【0096】画像からの人物位置検出処理の速さは、画像の入力レート、例えば、5フレーム/秒で行うので、音声処理をブロック単位で行う場合の処理速度とは一致せず音声処理より遅いのが普通である。従って、位置別発声動作情報検出部4-5に入力する画像は、音声処理に関する1ブロック前と同じものを使う場合があることになるが、その場合、同じ画像間の輝度変化を求めることになるので輝度変化は0になる。この状況を避けるため、輝度変化の計算の結果、値が0のときは1ブロック前の輝度変化の値をそのまま出力するようにする。

【0097】発声動作情報は、位置別到来パワーと並列に使うようにしており、環境適応部4-2の学習信号レベル計算部4-2cと、適応制御信号生成部4-2fにおいて用いている。本実施形態の他の部分は第2実施形

$$A_k = (P_N * 10^{v/10})^{1/2}$$

(位置別到来パワーが背景雑音+ v dBより小さいと ※ ※き)

$$A_k = (P_k)^{1/2}$$

(位置別到来パワーが背景雑音+ v dBより大きいとき)のように計算する。ここで、 P_N は背景雑音パワーの平均値、 P_k は位置別到来パワー、 v は最小値5である。

【0100】感度を低くするように設定した位置の場合 ★30

$$A_k = A_k * \gamma (P_N / P_k)^{1/2} * (1 - \gamma) (K_0 / (K_k + K_0))^{1/2} \quad (23)$$

によって計算する。ここで、 K_k は式(20)によって計算した k 番目の位置の顔周辺画像のフレーム間の輝度変化、 K_0 は同輝度変化の平均値、 γ は定数、例えば、 $\gamma = 0.5$ とする。以上のようにして求めた音源振幅と入力音声とを加算し、適応フィルタの入力信号を生成する。

【0101】次に、適応制御信号生成部4-2fは、位置別到来パワーと発声動作情報に基づいてフィルタ学習のステップサイズの制御を行う。ここでは、実施例2と☆

$$a = C / (\alpha_B P_k / P_N + \beta_B K_k + 1) \quad (\text{抑圧モードB}) \quad (24)$$

$$a = C / (\alpha_A \sum (P_k / P_N) + \beta_A \sum (K_k / K_0) + 1)$$

$$(\text{抑圧モードA}) \quad (25)$$

ここで、 P_k は感度を高くするように設定した位置 k からの位置別到来音パワー、 C 、 α_A 、 α_B 、 β_A 、 β_B は定数、例えば $C = 2.0$ 、 $\alpha_A = \alpha_B = 0.5$ 、 $\beta_A = \beta_B = 0.5$ である。ステップサイズの計算式として挙げた上式は一例であり、例えば、位置別到来パワーと輝度変化の値に対して各々しきい値を定め、どちらかー

*態と同じであり、環境適応部4-2の構成も同じであるので、学習信号レベル計算部4-2cと適応制御信号生成部6についてのみ述べる。

【0098】まず、学習信号レベル計算部4-2cでは、適応フィルタの入力信号と希望応答の生成の際の、到来を仮定する人物位置各々の音源信号の振幅を決定する。このため、第2実施形態と同様、学習信号レベル計算部4-2cでは、入力音声の背景雑音パワーと位置別到来パワーの音声区間の平均値を計算して保持し、上述の発声動作情報と、求めた位置別到来パワーと背景雑音パワーとから、音源信号の振幅を計算する。このとき、音源の存在を仮定する人物位置が、感度を高く設定した位置であるか、低く設定した位置であるかに応じて振幅の計算方法を変えるようにする。

【0099】感度を高くするように設定した位置の場合、第2実施形態と同様、位置別到来パワーの平均値の平方根の値を音源信号の振幅とするようにする。なお、発声がない場合は、位置別到来パワーは小さい値になってしまうため、背景雑音に対してある値 v 、例えば $v = 5$ dB高い値を最小値として設定し、位置別到来パワーがこれを上回る時だけ検出した値に設定するようにする。すなわち、

$$(21)$$

$$(22)$$

★合、位置別到来パワーと発声動作情報が大きいほど音源信号が小さい振幅となるようにして、人工的な学習信号への適応を弱めるようにする。例えば、 k 番目の音源信号の振幅を、

☆同様、この到来パワーの値が大きいときは適応を遅く、小さいときは適応を早くするようにステップサイズを制御する。

【0102】このため、例えば、式(5)で固定していたステップサイズ(式(5)の a)の値を次式により逐次計算して可変とすることにより、適応の速度を調整するようにする。

【0103】

方がこれを越えた場合に、適応を止める($a = 0$ とする)など、他の方法も使用可能である。

【0104】ここで、図15を参照しながら第3実施形態の環境適応部全体の処理の流れを説明する。

【0105】まず、環境適応部4-2の音源信号生成部4-2bにおいて、選択人物位置の数の系列の音源信号

を生成する（ステップS21）。

【0106】次に、学習信号レベル計算部4-2cにおいて、複数チャネルで入力した入力音声のパワーを計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーから背景雑音パワー P_N を求める。このとき、式（12）により平均化してもよい（ステップS22）。

【0107】次に、複数チャネルの入力音声から式（15）により位置別到来パワーを計算する（ステップS23）。次に、位置別発声動作情報検出部4-5において、人物位置ごとの顔周辺画像を人物位置検出部2から入力し、発声動作情報を検出する。輝度変化の計算の結果、値が0ならば1ブロック前の値をこのブロックの輝度変化の値とし、0以外なら計算結果をこのブロックの輝度変化の値とし、この値を記憶する（ステップS24）。

【0108】次に、学習信号レベル計算部4-2cにおいて、式（21）から（23）により音源信号の振幅 A_k を計算する（ステップS25）。次に、入力信号生成部4-2aにおいて、式（13）により、実際の入力音声と加算して適応フィルタの入力信号を生成する（ステップS26）。次に、希望応答生成部4-2dにおいて、式（14）により、音源信号と音源信号の振幅から希望応答を生成する（ステップS27）。

【0109】次に、適応制御信号生成部4-2fにおいて、式（24）または（25）により、背景雑音パワーと位置別到来パワーと位置別発声動作情報とからステップサイズの系列である適応制御信号を生成する（ステップS28）。次に、生成した入力信号と希望応答と適応制御信号を適応フィルタに入力し、フィルタ係数を得る（ステップS29）。得られたフィルタは、雑音除去部4-3に送り、入力音声とフィルタを畳み込んで音声抽出する。

【0110】第3実施形態の全体の処理の流れは第1実施形態と同じであるので改めて述べない。以上に述べたように、画像処理により特定された複数の人物位置各々から到来する音声に対し、適応フィルタにより感度を設定して雑音抑圧を行う際、人物位置からの到来音のパワーと画像から求めた発声動作情報に応じて適応フィルタの適応処理を制御しているため、雑音が大きく、位置別到来パワーの推定が低い場合でも、雑音抑圧性能を大幅に高くしながら、複数の人物が同時に発声した場合に、その人物すべての音声を背景雑音を抑圧して取り出したり、一人だけの人物の音声のみを他の人物の音声を抑圧して取り出したりできる。

【0111】以下にテスト発声モードを備えた第4実施形態について説明する。第4実施形態は、音声収集装置の動作中に、一時、通常の音声入力処理を停止し、テスト発声を入力してレベル計算を行うための、入力モード切り替え部を具備することにより、音源信号のレベルを

実環境の値に合わせ、高精度の適応処理を行うものである。

【0112】これまでに述べた実施形態では、適応フィルタの学習信号の生成に用いる音源信号は、背景雑音レベルと経験的に決めたデフォルトの音声のレベル値を用いてその振幅を計算してきたが、本実施形態では、さらに現実の音場に正確に合わせるため、テスト発声を行って音源のレベルを決めるようにしている。このため、テスト発声か、通常の音声入力かを動作中に切り替える入力モード切り替え部7を追加し、図16に示すような構成としている。図において、1は画像入力部、2は人物位置検出部、3は音声入力部、4は雑音抑圧部、5は入力モード切り替え部である。

【0113】この構成において、通常は入力モード切替部には、通常の音声入力処理であることを設定しておき、テスト発声時には入力モード切替部から、テスト発声であることを入力する。入力モードをテスト発声に設定した場合、通常行っている適応フィルタ処理は止め、学習信号レベル計算部4-2cにおいて、入力音声のレベルを計算し、保持するようにする。テスト発声を終了して通常の入力モードに戻った際は、学習信号レベル計算の際、デフォルトで例えば5dBなどと決めてきた音源信号の最小値を使わず、ここで求めたテスト発声のレベルから音源信号の振幅を計算するようにする。

【0114】ここで、図17を参照しながら第4実施形態の全体の処理の流れを説明する。

【0115】まず、初期設定を行い、選択する人物位置の数Nと雑音抑圧処理のモードAかBかを設定する（ステップS41）。

【0116】画像の処理の方では、画像データをビデオカメラから、例えば毎秒5フレームで取り込み（ステップS42）、フレームごとに人物位置を特定して出力し（ステップS43）、このステップS42とS43を繰り返す。画像から人物位置を特定する画像の処理は、音声の雑音抑圧処理とは独立に、並列に処理するようにする。

【0117】音声処理の方では、まず、音声データを、例えばサンプリング12kHzでAD変換し、1チャネルあたり、例えば1024サンプルを1ブロックとして1ブロック分のデータを取り込む（ステップS44）。

【0118】次に、ステップS45で入力モードがテスト発声か通常入力かを検査し、テスト発声であればステップS46に進み、通常入力であればステップS47に進む。ステップS46では、学習信号レベル計算部4-2cにおいて、入力音声のレベルを計算して保持する。入力音声のレベルは、ある番号、例えば1ch目のマイクホンからの入力のパワーに基づいて音声検出を行い、音声区間として検出された部分の平均値を用いるようにする。この後、ステップS44に戻る。

【0119】次に、ステップS47では人物位置が特定

されているか否かを判定し、人物位置が特定されていない場合は、何もせずにこのステップS44乃至S47を繰り返し、特定された場合は次のステップS48に進む。人物位置は、処理開始直後で画像処理結果が出ていない場合や人物がいない場合に特定されない。位置画像に関する処理と音声に関する処理とは独立しているため、例えば、一つの計算機上で全処理を行う場合、よく知られているように、ソケットを用いたプロセス間通信やシェアドメモリ、あるいはファイルを通じて人物位置のデータのやり取りを行うことができる。

【0120】次のステップS48では、人物位置選択部4-1において、処理対象とする人物位置を選択する。次に、環境適応部4-2において、人物位置選択部4-1で選択された人物位置または方向と距離を用いて適応フィルタの学習信号を生成し、フィルタ係数を更新する(ステップS49)。次に、ステップS49で更新されたフィルタ係数を雑音除去部4-3にコピーし、このフィルタと入力音声との畳み込み演算を行って音声を出力する(ステップS50)。

【0121】以上のステップS41からS42の処理とステップS43からS50までの処理を並列に繰り返す。

【0122】なお、本実施例で述べたテスト発声モードは、第1実施形態に追加する形で述べたが、第2、第3実施形態に述べた構成に追加して併用することも可能である。

【0123】次に、本発明の第5実施形態について説明する。本実施形態では、人物ごとの音声を他の人物の音声と分離して取り出すため、第1乃至第4実施形態の雑音除去部と、環境適応部における適応処理部各々を複数のフィルタから構成するようにしている。これを、図18に示す。この部分以外については第1乃至第4実施形態と同様の構成である。なお、この実施形態は、第3実施形態の拡張として説明するが、第2、第4実施形態にも適用可能であり、また、環境適応部の適応制御信号生成部を取り去るだけで第1実施形態を拡張した場合にもなっている。図18において、環境適応部4-2における適応処理部4-2eと雑音除去部4-3におけるフィルタは各々複数個(N個)づつ、例えば、3個づつ用意し、雑音除去部4-3のフィルタは環境適応部4-2の適応フィルタの係数のコピーである。また、環境適応部4-2の適応フィルタには、すべて同じ入力信号を入力するが、希望応答と適応制御信号は、フィルタの番号kにより異なった内容のものを入力するようにする。

【0124】この適応処理部4-2eを含む環境適応部4-2の処理について、次に説明する。まず、環境適応部4-2では、画像により検出され選択された人物位置各々から音波が到来すると仮定し、その音波の信号を音源信号生成部4-2bで生成する。この信号の内容は、人工的なもの、例えば音源間で無相関なランダム雑音で

良いことは、第1実施形態で述べた。この音源信号をもとに、適応処理を行うための入力信号と希望応答を生成する。その際、学習信号レベル計算部4-2cにおいて、位置別到来パワー、発声動作情報のいずれか又は両方と観測した背景雑音レベルとに基づいて音源信号の振幅を決定する。

【0125】また、適応処理の際の適応速度の制御を行う適応制御信号を適応制御信号生成部4-2fにおいて生成する。適応処理部4-2eでは、上記3つの信号を入力として適応フィルタにより雑音抑圧のためのフィルタ係数を決定する。なお、適応制御信号は必ずしも必要でなく、また、学習信号レベルの計算には、位置別到来パワーと発声動作情報は必ずしも必要でない。

【0126】人物位置検出部2により複数検出され、その中から人物位置選択部4-1により選択された複数の人物位置各々から到来する音声の抽出を、構成図で示したように、複数個のフィルタを使って行う。フィルタの数は選択人物位置の数と一致するようにし、フィルタの番号kは人物位置の番号に対応させるようにする。

【0127】各フィルタが各人物の音声を抽出するようにするため、k番目の適応フィルタによる適応の際の希望応答の内容は、k番目の人物位置から到来することを仮定する1個の音源の信号と同じになるようにし、式

(14)により計算される。また、各適応フィルタの入力信号は、N個の人物位置から各々に対応する音源信号が伝搬してマイクロホン位置で観測されるときに信号をすべて重ね合わせたものに実際に入力した音声を加算したものであり、式(13)により、マイクロホン位置ごとに計算されてNチャンネルの信号が生成される。入力信号はすべての適応フィルタで共通に使われる。

【0128】一方、適応フィルタの収束速度を制御する適応制御信号は、適応フィルタの番号ごとに異なった信号内容のものを生成するようにし、k番目のフィルタには、k番目の人物位置からの位置別到来パワーまたは発声動作情報に基づいて式(19)、(20)または式(24)、(25)により計算されたステップサイズの値の系列を入力するようにする。位置別到来パワーまたは発声動作情報が得られない第1実施形態を拡張する場合は、適応制御信号は生成せず、ステップサイズは一定値とする。

【0129】上記のようにして生成した入力信号、希望応答、適応制御信号を適応フィルタに入力して複数組のフィルタ係数を決定した後、これらの係数を雑音除去部4-3に送り、入力音声をフィルタ処理して雑音を除去し、各人物の音声を別々に抽出するk番目の人物位置の音声はk番目のフィルタから出力されることになる。

【0130】以上のように、人物位置に対応した複数のフィルタを用いることにより、各人物位置からの到来音を別々に分離して取り出すことが可能となる。

【0131】次に、本発明の第6実施形態について説明

する。第6実施形態はパイロット信号法による適応フィルタでなく、線形拘束条件付き適応フィルタにより、雑音抑圧処理を行って音声収集するものである。この種類の適応フィルタにより計算量の多い学習信号の生成処理を省いた処理が可能である。

【0132】図19は第6実施形態の全体構成を示す図である。図19において、1は画像を入力する画像入力部、2は入力した画像を処理して人物の位置を特定する人物位置検出部、3は複数のマイクロホンからの音声と並列に入力する音声入力部、4は画像処理により検出された人物位置に基づいて複数のマイクロホンからの入力音声から雑音を抑圧して音声を取り出す雑音抑圧部である。この雑音抑圧部4の内部構成は、人物位置検出部2により特定された人物位置の中から処理対象とする人物位置を選択する人物位置選択部4-1と、選択された人物位置に基づいて適応フィルタの拘束条件の設定を行う拘束条件設定部4-2と、設定された拘束条件のもとで適応フィルタにより雑音抑圧処理を行う雑音除去部4-*

$$E[y^2] = E[w^H X X^H w] = w^H R w \quad (E[\] \text{は期待値})$$

(26)

を、目的の方向または位置に対する応答を一定に保つという拘束条件下で最小にすることにより得られる。ここ※

$$W^H A = g$$

と表される。ここで、 g は拘束条件の数 G の大きさの定数値の列ベクトルで、例えば $[1, 1, \dots, 1]$ であ

$$A = [a_1, \dots, a_L]$$

と表される。上式(6)の成分の各方向制御ベクトル a_m ($m=1, \dots, L$)は

【数1】

$$a_m = (1, a_2 e^{-j\omega_m \tau_2}, \dots, a_N e^{-j\omega_m \tau_N}) \quad (28)$$

【0136】である。ここで、 τ_2, \dots, τ_N は1番目のマイクロホンを基準としたときの各マイクロホンに入射する音波の伝搬時間差、 ω_m は角周波数、 a_2, \dots, a_N は1番目のマイクロホンを基準としたときの各マイクロホンに入射する音波の振幅比である。 G は例えば10を用い、 ω_m は例えば $\omega_m = ((\omega_a - \omega_b) / (G \star$

$$\begin{aligned} a_m(\theta_1) &= (1, a_2 e^{-j\omega_m \tau_2(\theta_1)}, a_3 e^{-j\omega_m \tau_3(\theta_1)}, \dots, \\ a_m(\theta_2) &= (1, a_2 e^{-j\omega_m \tau_2(\theta_2)}, a_3 e^{-j\omega_m \tau_3(\theta_2)}, \dots, \end{aligned}$$

*3とからなる。

【0133】人物位置選択部4-1では、第1実施形態において述べたように、画像から得られた複数の人物位置から定めた数の人物位置の選択を行い、拘束条件設定部4-2ではこの人物位置に基づき、線形拘束条件付き適応フィルタの拘束条件を設定する。拘束条件によって、任意の人物位置から到来する音波に対する感度を設定できるようになる。雑音除去部4-3では、設定された拘束条件のもとで適応フィルタにより雑音抑圧処理を行う。

【0134】線形拘束条件付き適応フィルタの詳細は、例えば、文献(Heykin著: Adaptive Filter Theory)に詳述されているが、一応、処理方法を述べる。

【0135】式(1)から(4)を参照し、マイクロホンアレイの出力を X 、フィルタ係数を W 、フィルタの出力を $y = W^H X$ とすると、拘束条件付き最小分散適応フィルタのフィルタ係数は、次式によるフィルタの出力パワー y^2 の期待値

※で、 $R = E[X X^H]$ は X の自己相関行列である。また、拘束条件は、

$$(26-1)$$

★り、 A は異なった周波数に関する方向制御ベクトル a_m を列ベクトルとする行列であり、

$$(27)$$

☆-1) * $m + \omega_b$ とする。ここで ω_a は帯域の上限、 ω_b は下限の角周波数である。

【0137】式(26-1)の拘束条件として、一つの方向または位置から到来する音波に関する応答を一定にするだけでなく、複数の方向または位置から到来する音波に対する応答を同時に一定にするようにする。例えば、 θ_1, θ_2 の2つの到来角度に関する時間遅れ $\tau_i(\theta_1), \tau_i(\theta_2)$ (式(6))を用いた方向制御ベクトル $a_m(\theta_1), a_m(\theta_2)$ ($m=0, 1, \dots, L$)、

【数2】

$$a_{N-1} e^{-j\omega_m \tau_{(N-1)}(\theta_1)} \quad (29)$$

$$a_{N-1} e^{-j\omega_m \tau_{(N-1)}(\theta_2)} \quad (30)$$

【0138】を用いて、

$$A = [a_0(\theta_1), a_1(\theta_1), \dots, a_L(\theta_1), a_0(\theta_2), a_1(\theta_2), \dots, a_L(\theta_2)] \quad (30)$$

とすることで、複数の到来方向に対するアレイの応答の拘束条件を設定することができる。

【0139】ここで、式(4)と(5)による最小化問

題を反復的に求める場合、 j 回めの反復による更新後のフィルタ係数は、次式のように表される。

【0140】

$$W_j = P [W_{j-1} - \mu y_j X] + F$$

ここで、PとFは、

$$P = I - A (A^H A)^{-1} A^H, F = A (A^H A)^{-1} g \quad (32)$$

である。式(8)により、雑音を抑圧して目的の音声を取り出すフィルタ係数が得られるとともに、雑音を抑圧した音声出力 y_j が同時に得られることになる。次に、雑音抑圧のための拘束条件の設定について説明する。複数の人物位置が得られた場合の雑音抑圧の仕方は、第1実施形態で述べたように、処理対象として選択されたすべての人物位置からの到来音波を高い感度で得るようにする抑圧処理モードAと、選択された人物位置の中の一つから到来する音波のみ高い感度にし他の人物位置からの音波に対しては感度を低くする抑圧処理モードBがある。他にも、AとBの中間の方法として、所定の複数の人物位置に対して感度を高くしその他に対しては低くするなどが考えられるが、AとBの組み合わせで実現できる。

【0141】雑音抑圧の拘束条件は、拘束条件を表す式(25)において、行列Aの要素と定数ベクトル g を与えることにより設定する。処理モードAもBも、行列Aの内容は同じであり、選択した人物位置に関する方向制御ベクトル式(30)である。定数ベクトル g の内容は抑圧処理モードに応じて変えるようにし、選択した人物位置すべてに対して感度を高くする抑圧モードAの場合、 g の要素はすべて1とし、抑圧処理モードBの場合、高い感度に設定する人物位置に関する g の要素は1とし、低い感度に設定する人物位置に関する g の要素は0にする。

【0142】例えば、方向 θ_1 、 θ_2 に関する方向制御ベクトルの行列Aが、次に示す式(30)の内容の場合、

$$A = [a_0(\theta_1), a_1(\theta_1), \dots, a_L(\theta_1), a_0(\theta_2), a_1(\theta_2), \dots, a_L(\theta_2)]$$

方向 θ_1 に対して感度を高くし、 θ_2 に対して感度を低くする場合の定数ベクトル g の内容は、

$$g = [1, 1, \dots, 1, 0, 0, \dots, 0]$$

とする。

【0143】次に、以上の処理の流れについて図20を参照しながら説明する。

【0144】第1実施形態で述べたように、画像から人物位置を特定する画像の処理と、音声の雑音抑圧処理とは、並列に処理するようにし、画像処理の方は第1実施形態と同じである。

【0145】まず、初期設定を行い、選択する人物位置の数Nと雑音抑圧処理のモードAかBかを設定する(ステップS51)。

【0146】画像の処理の方では、画像データを、例えば毎秒5フレームで取り込み(ステップS52)、フレームごとに人物位置を特定する(ステップS53)。

$$(31)$$

【0147】音声処理の方では、まず、音声データを例えばサンプリング12kHz、1チャンネルあたり1024サンプルを1ブロックとして1ブロック分取り込む(ステップS54)。

【0148】次に、人物位置が特定されているか否かを判定し(ステップS55)、人物位置が特定されていない場合は、ステップS54に戻り、特定されている場合は次のステップS56に進む。

【0149】次のステップS56では、人物位置選択部4-1において、処理対象とする人物位置を選択する。次に、選択された人物位置に基づいてフィルタ処理の拘束条件を式(26)、(30)に従って設定する(ステップS57)。

【0150】次に、ステップS57で設定した拘束条件のもとに適応フィルタの演算を行って音声出力する(ステップS58)。以上ステップS52からS53の処理とS54からS58までの処理とを並列に繰り返す。

【0151】以下に、拘束条件付きと複数のフィルタを備えた第7実施形態を詳細に説明する。第7実施形態は、拘束条件付き適応フィルタを使った場合に、複数の人物位置各々からの到来音を分離して取り出すものである。複数の人物位置からの到来音を分離して取り出すため、全体構成図の雑音除去部4-3を複数の適応フィルタにより図21に示すように構成する。

【0152】図21において、適応フィルタの数は人物位置選択部4-1において選択する人物位置の数と一致させてN個、例えば3とし、適応フィルタごとに異なった内容で拘束条件を設定する。拘束条件は、拘束条件設定部4-2において行い、k番目の適応フィルタには、選択された人物位置の中のk番目の位置に対して感度を高くし、他の人物位置に対しては感度を低くするように設定した拘束条件を入力する。

【0153】以上のように、複数の適応フィルタを用い、各々に異なった拘束条件を設定することにより、人物位置ごとの到来音を他の位置からの到来音と分離して抽出することが可能となる。

【0154】以下に、拘束条件付き適応フィルタとパイロット信号法を組み合わせた第8実施形態を説明する。第8実施形態の構成を図22に示す。図22において、1は画像入力部、2は人物位置検出部、3は音声入力部、4は雑音抑圧部であり、雑音抑制部4の内部は、人物位置選択部4-1と、環境適応部4-2と、雑音除去部4-3と、拘束条件設定部4-5とからなる。

【0155】この構成において、画像から検出して人物位置の中から人物位置選択部4-1により複数を選択し、これに基づいて拘束条件設定部4-5において線形

拘束条件付き適応フィルタの拘束条件を設定し、環境適応部4-2においてこの適応フィルタの学習信号である入力信号と希望応答を生成して適応フィルタに入力して雑音抑圧のためのフィルタ係数を決定し、決定したフィルタ係数を雑音除去部4-3に送り、入力音声をフィルタ処理して雑音を除去する。

【0156】拘束条件と学習信号の作成方法は種々考えられるが、ここでは、抑圧処理のモードBの場合について説明する。この場合、一つの人物位置に対して感度を高くなるように拘束条件を設定し、他の人物位置に対しては感度が低くなるように学習信号を生成して適応フィルタの係数を決定する。

【0157】例えば、方向 θ_1 に関して感度を高く設定する場合、方向制御ベクトルの行列Aを、次に示す内容にし、

$$A = [a_0(\theta_1), a_1(\theta_1), \dots, a_L(\theta_1)]$$

定数ベクトルgの内容は、

$$g = [1, 1, \dots, 1]$$

とする。

【0158】また、学習信号のうち、入力信号は、感度を低く設定する人物位置からのみの音波の到来を仮定し、マイクロホン位置で観測される信号を第1実施形態の式(13)により求める。この場合、希望応答は使わないので生成しない。従って、環境適応部は、第1、第2実施形態の環境適応部における希望応答生成部を除いた図23に示すような構成により実現可能である。

【0159】図23において、4-2cは学習信号レベル計算部、4-2aは入力信号生成部、4-2bは音源信号生成部、4-2eは適応処理部である。この構成により、拘束条件付き適応フィルタの拘束条件を設定した後、入力信号を適応フィルタに入力してフィルタ係数の更新を行う。

【0160】上述の環境適応部4-2の処理の流れを図24を参照しながら説明する。

【0161】まず、環境適応部4-2の音源信号生成部4-2bにおいて、選択人物位置の数の系列の音源信号を生成する(ステップS61)。

【0162】次に、学習信号レベル計算部4-2cにおいて、複数チャンネルで入力した入力音声のパワーを計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーから背景雑音パワー P_N を求める。このとき、式(12)により平均化してもよい(ステップS62)。

【0163】次に、学習信号レベル計算部4-2cにおいて、式(11)により音源信号の振幅 A_k を計算する(ステップS63)。次に、入力信号生成部4-2aにおいて、式(13)により、実際の入力音声と加算して適応フィルタの入力信号を生成する(ステップS64)。次に、生成した入力信号を適応制御信号として適

応フィルタに入力し、フィルタ係数を得る(ステップS65)。

【0164】得られたフィルタは、雑音除去部4-3に送り、入力音声とフィルタを畳み込んで音声を抽出する。雑音除去部4-3におけるフィルタ処理は、式(4)に従って行う。

【0165】次に、本実施例全体の処理の流れについて、図25を参照して説明する。

【0166】第1実施形態で述べたように、画像から人物位置を特定する画像の処理と、音声の雑音抑圧処理とは、並列に処理するようにし、画像処理の方は第1実施形態と同じである。

【0167】まず、初期設定を行い、選択する人物位置の数Nと雑音抑圧処理のモードAかBかを設定する(ステップS71)。

【0168】画像の処理の方では、画像データを、例えば毎秒5フレームで取り込み(ステップS72)、フレームごとに人物位置を特定する(ステップS73)。

【0169】音声処理の方では、まず、音声データを例えばサンプリング12kHz、1チャンネルあたり1024サンプルを1ブロックとして1ブロック分取り込む(ステップS74)。

【0170】次に、人物位置が特定されているか否かを判定し(ステップS75)、人物位置が特定されていない場合は、ステップS74に戻り、特定されている場合は次のステップS76に進む。

【0171】次のステップS76では、人物位置選択部において、処理対象とする人物位置を選択する。

【0172】次に、選択された人物位置に基づいてフィルタ処理の拘束条件を式(26)、(30)に従って設定する(ステップS77)。

【0173】次に、環境適応部4-2において、人物位置選択部4-1で選択された人物位置または方向と距離を用いて適応フィルタの学習信号を生成する(ステップS78)。

【0174】ステップS77で設定した拘束条件のもとに適応フィルタの演算を行ってフィルタ係数を更新し、雑音除去部4-3にフィルタ係数を転送する(ステップS79)。

【0175】次に、雑音除去部4-3において、ステップS79で転送されたフィルタと入力音声との畳み込み演算を行って音声を出力する(ステップS80)。

【0176】以上のステップS72からS73の処理と、S74からS80までの処理を並列に繰り返す。

【0177】以上に述べたように、拘束条件付き適応フィルタにパイロット信号法を適用することにより、パイロット信号法で必要な学習信号生成のための処理量を減らし、また、拘束条件付き適応フィルタにおいて、拘束条件が多い場合のフィルタの自由度低下による性能低下を避けることができるため、少ない処理量で精度よく実

環境に適応して雑音抑圧処理を行うことができる。

【0178】以下に第9実施形態を説明する。第9実施形態は、拘束条件付き適応フィルタとパイロット信号法を組み合わせた場合に、複数の人物位置からの音声を分離して取り出すものである。複数の人物位置からの音声を分離して取り出すため、雑音除去部4-3と環境適応部4-2の適応処理部4-2eとを、図26に示すように複数のフィルタから構成するようにしている。

【0179】図26において、適応処理部4-2eの適応フィルタと雑音除去部4-2のフィルタはN個、例えば3個ずつ用意し、適応処理部4-2eで決定した適応フィルタの係数を雑音除去部4-3に送るようにしている。適応処理部4-2eの各適応フィルタの入力には、入力信号生成部4-2aで生成した入力信号を共通に入力し、拘束条件は適応フィルタごとに異なったものを入力する。

【0180】拘束条件は拘束条件設定部4-5において設定し、k番目の適応フィルタには、人物位置選択部4-1において選択されたk番目の人物位置から到来する音波に対して感度を高くなるようにした拘束条件が入力される。拘束条件の設定方法は第8実施形態で述べたのと同じであり、また、適応フィルタの入力信号の生成方法も同じである。

【0181】以上に述べたように、拘束条件付き適応フィルタとパイロット信号法を組み合わせた場合に、複数のフィルタにより適応処理を行っているため、人物位置ごとの到来音を分離して抽出することができ、且つ、パイロット信号法で必要な学習信号生成のための処理量を減らし、また、拘束条件付き適応フィルタにおいて、拘束条件が多い場合のフィルタの自由度低下による性能低下を避けることができるため、少ない処理量で精度よく実環境に適応して雑音抑圧処理を行うことができる。

【0182】なお、上記した人物位置決定工程と、人物位置選択工程と、フィルタ係数決定工程と、音声抽出工程とはコンピュータプログラムとして、ハードディスク、フロッピーディスク、CD-ROMなどの記憶媒体に記憶し、この記憶媒体を適当な計算機に搭載して実行することができる。

【0183】

【発明の効果】本発明によれば、複数の人物位置からの音声に対して、背景雑音を抑えてすべての音声を同時に抽出するかあるいは、特定の人物位置からの音声のみを抽出することができる。

【図面の簡単な説明】

【図1】本発明の第1実施形態に係る音声収集装置の構

成を示す図である。

【図2】カメラとマイクの配置を示す図である。

【図3】画面上の人物データの一例を示す図である。

【図4】人物位置データの一例を示す図である。

【図5】環境適応部の構成を示す図である。

【図6】フィルタの構成を示す図である。

【図7】マイクロホンと人物位置の設定を示す図である。

【図8】環境適応部の処理の流れを示すフローチャートである。

【図9】第1実施形態全体の処理の流れを示すフローチャートである。

【図10】第2実施形態における雑音抑圧部の構成を示す図である。

【図11】環境適応部の構成を示す図である。

【図12】位置別到来パワーに基づく学習信号の生成について説明するための図である。

【図13】環境適応部の処理の流れを示すフローチャートである。

【図14】第3実施形態の全体構成を示す図である。

【図15】環境適応部の処理の流れを示すフローチャートである。

【図16】第4実施形態の全体構成を示す図である。

【図17】第4実施形態の全体処理の流れを示すフローチャートである。

【図18】第5実施形態における雑音除去部と環境適応部の構成を示す図である。

【図19】第6実施形態の全体構成を示す図である。

【図20】第6実施形態の処理の流れを示すフローチャートである。

【図21】第7実施形態の雑音除去部の構成を示す図である。

【図22】第8実施形態の全体構成を示す図である。

【図23】第8実施形態における環境適応部の構成を示す図である。

【図24】環境適応部の処理の流れを示すフローチャートである。

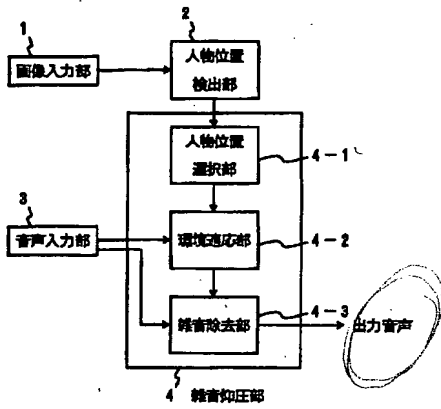
【図25】第8実施形態の処理の流れを示すフローチャートである。

【図26】第9実施形態における雑音除去部と環境適応部の構成を示す図である。

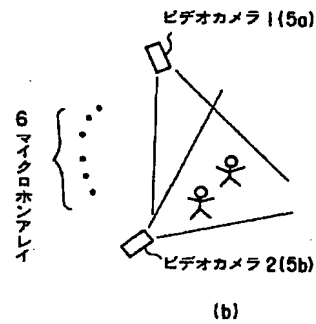
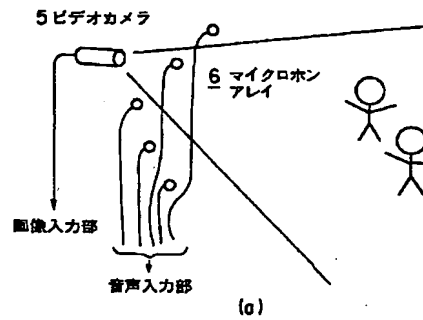
【符号の説明】

1…画像入力部、2…人物位置検出部、3…音声入力部、4…雑音抑圧部、4-1…人物位置検出部、4-2…環境適応部、4-3…雑音除去部。

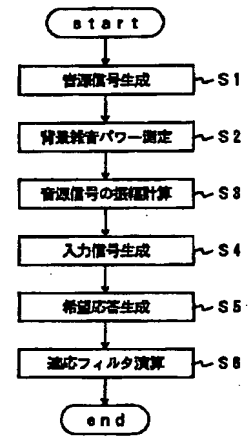
【図1】



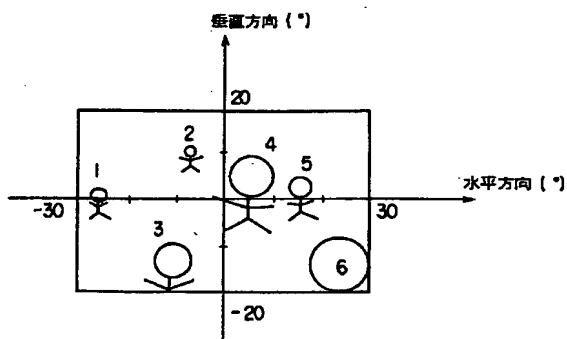
【図2】



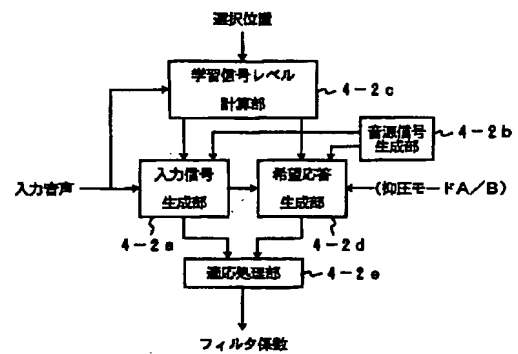
【図8】



【図3】



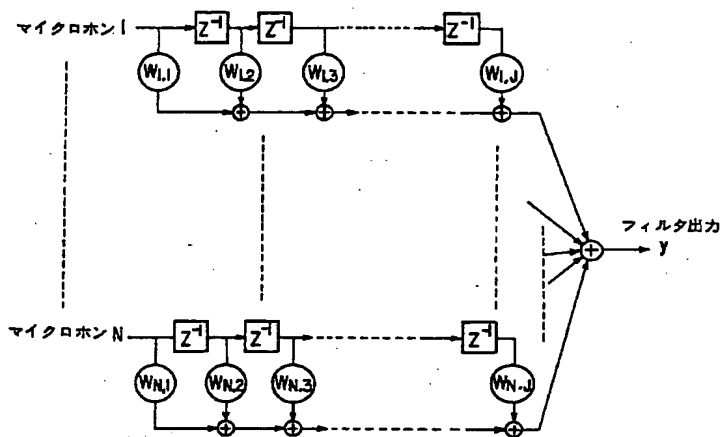
【図5】



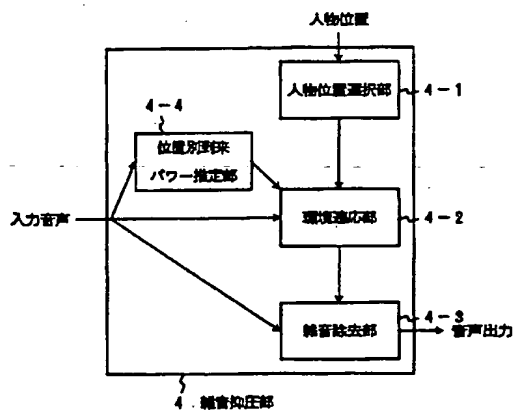
【図4】

番号	人物方向		A 人物の顔部分の 面積 (画素数)	B カメラ中心線方向と人物方向差 ($\sqrt{X^2+Y^2}$)	A/B
	水平 (X)	垂直 (Y)			
1	-25°	0°	80	25	3.2
2	-7°	10°	65	12.2	5.3
3	-10°	-12°	300	15.6	19.2
4	5°	5°	400	7.1	56.3
5	15°	3°	120	15.8	7.6
6	25°	-15°	500	28.2	17.1

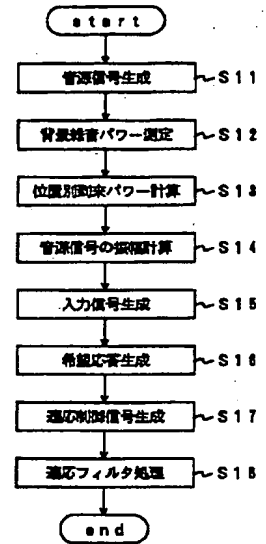
【図6】



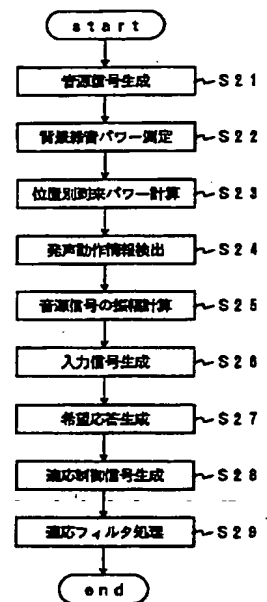
【図10】



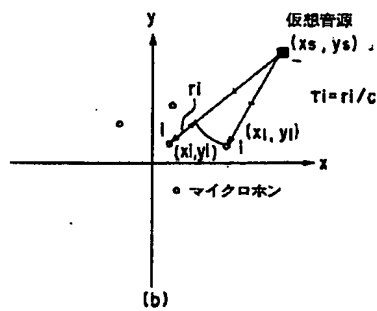
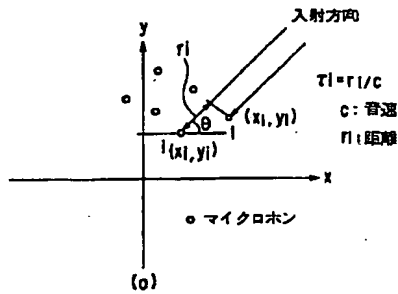
【図13】



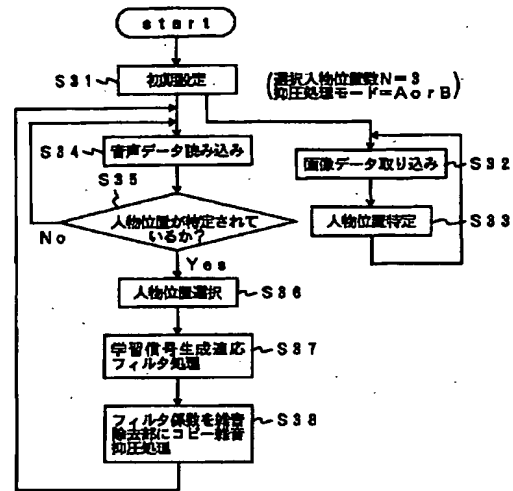
【図15】



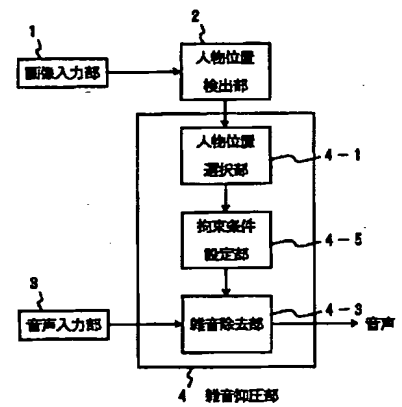
【図7】



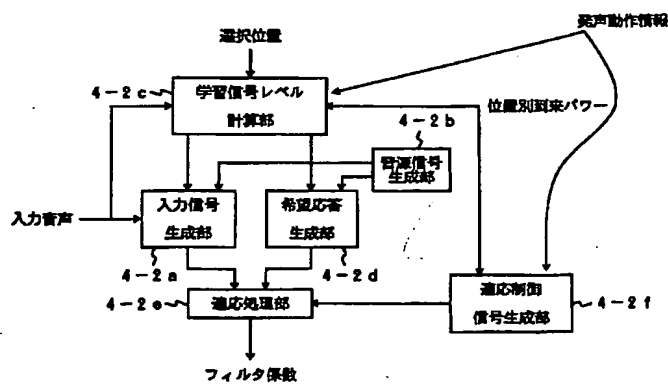
【図9】



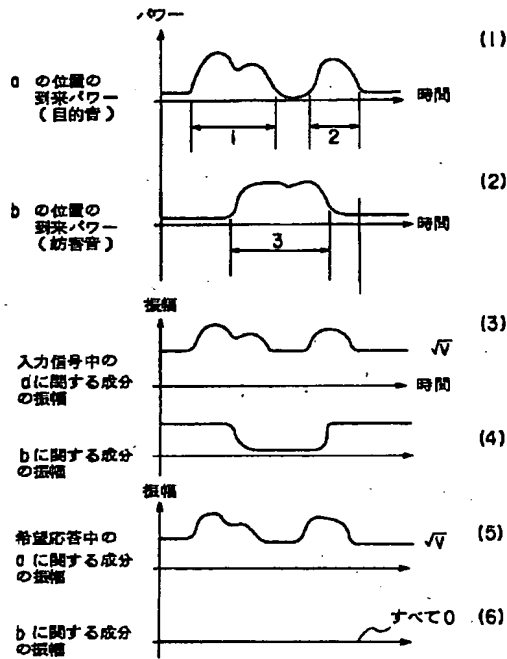
【图 19】



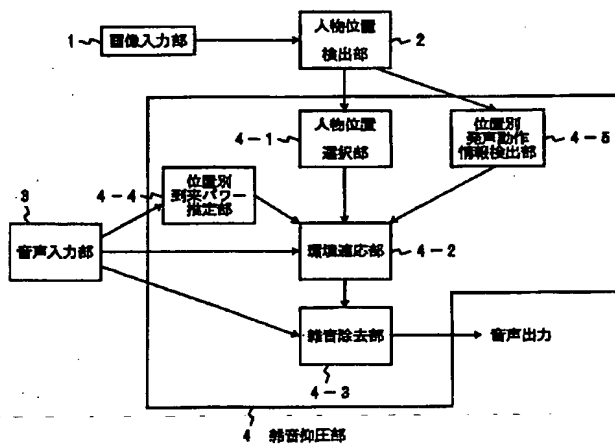
【図 1 1】



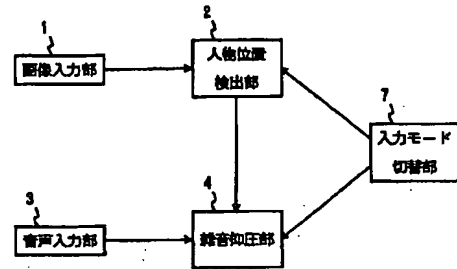
【図12】



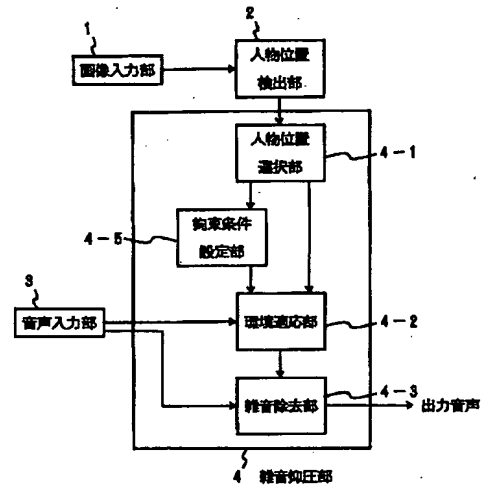
【図14】



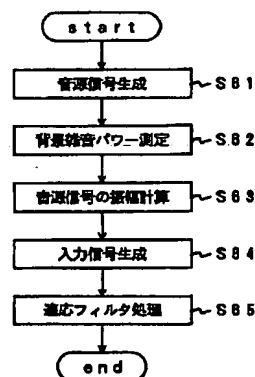
【図16】



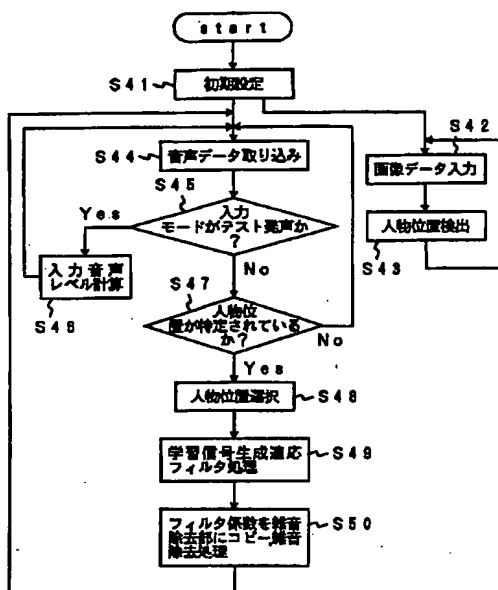
【図22】



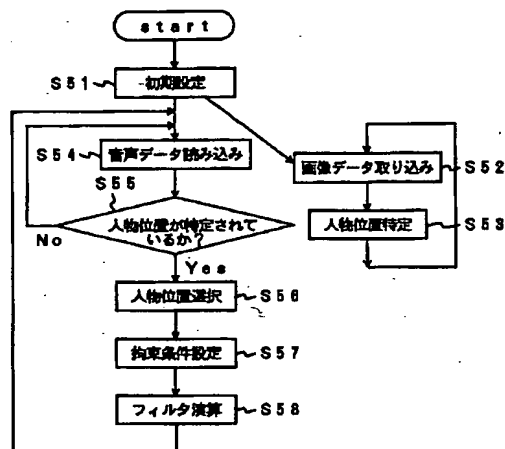
【図24】



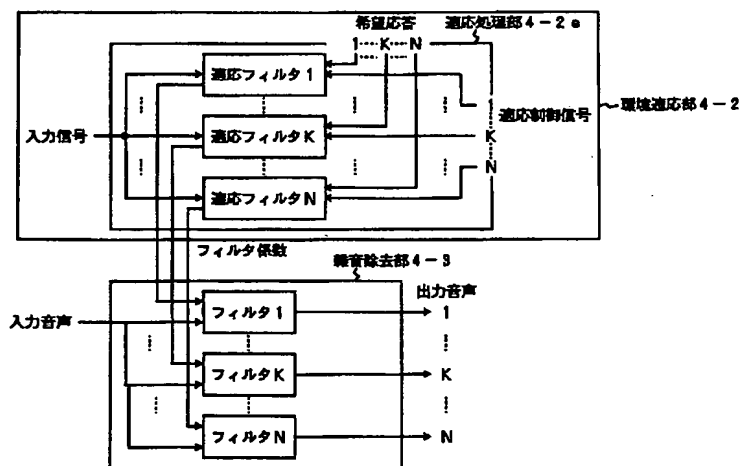
【図17】



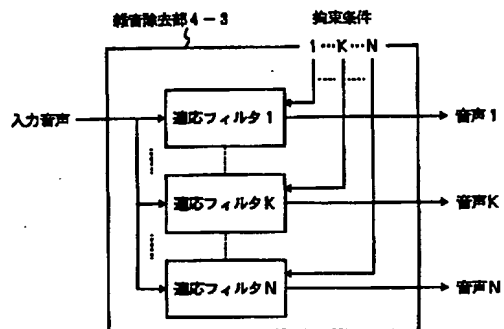
【図20】



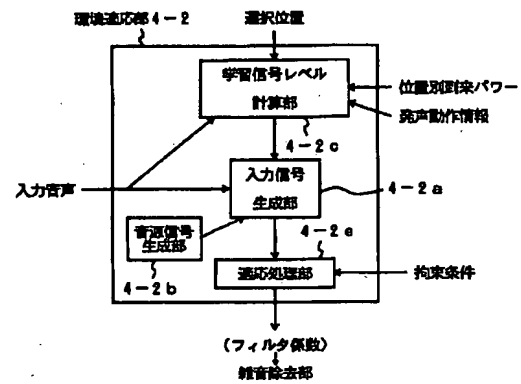
【図18】



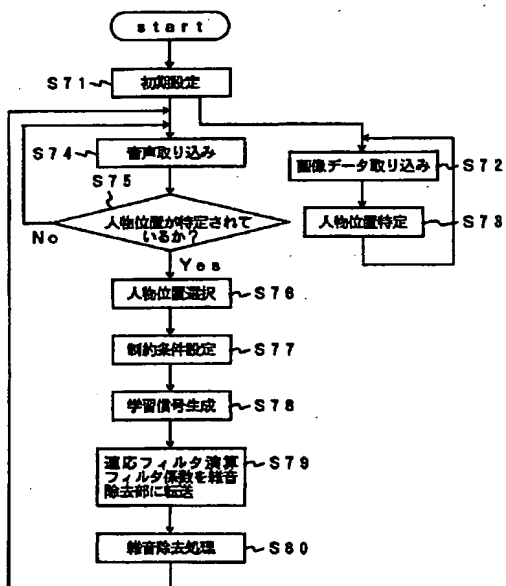
【図21】



【図23】



【図25】



【図26】

